

Hard battles on radiation safety

The International Commission on Radiological Protection has been the unaccustomed target of public criticism for several months. The critics are wrong, but the commission needs to mend its ways.

THE International Commission on Radiological Protection (ICRP) is an unusual animal, but that does not explain why so many brickbats have been thrown at it in the past few weeks, running up to its biannual meeting, finishing this week at Como, Italy. All kinds of people have been demanding that ICRP should this year revise its recommendations about the doses of radiation to which people should be allowed to be exposed, making them more stringent. Three hundred scientists have signed a petition to that effect (see *Nature* 329, 239; 1987). The Friends of the Earth have put out press releases. A correspondent has written in the same vein to *Nature* asking that his letter should be published not later than this week so that ICRP will do as asked. (Unfortunately that is not possible.) The enormously successful and generally respected weekly newspaper *The Economist* has carried a stirring leading article demanding that ICRP should forthwith recommend tighter limits for public exposure to radiation. What ICRP's sudden flock of critics appears not to appreciate is that these demands cannot be met without undermining the foundations of ICRP's credibility.

ICRP began life in 1928 as a kind of radiologists' self-help committee, concerned to pool information about the untoward exposure of professional people to radiation and to devise acceptable standards for their protection. For most of its existence, the commission has been run by volunteers out of back rooms, but others have latterly stepped in to help. Throughout, ICRP has been jealous of its independence from governments, although it now receives some financial help from United Nations agencies. It sprang to prominence only in the 1950s, when bomb-testing governments and those busy developing civil nuclear power stations found themselves in need of acceptable radiation standards and discovered, much to their surprise, that ICRP had already done much of the work. At that stage (and until 1960), *The Economist* now complains, ICRP believed that there was a threshold dose of radiation below which exposure would not bring untoward effects, cancer at some later stage, for example. (So, too, did *The Economist*, as it happens.)

Limits

ICRP's current recommendation on the exposure of members of the public to artificial sources of radiation, promulgated in 1955 after a meeting in Paris, is that the average radiation dose received by a member of the public should be limited to "a lifetime average annual dose of 1 mSv", with the proviso that groups of people exposed to as much as 5 mSv in a year should be dealt with so that their accumulated lifetime dose will not exceed the allowed 1 mSv a year. In this quotation, "dose" should strictly be preceded by "effective" to allow for the fact that some kinds of ionizing radiation (neutrons, for example) are more damaging than others. In the old units, 1 mSv is 100 mrem. By comparison, the average dose to which people are exposed by natural causes is of the order of 200 mrem a year, allowing for internal sources of radiation such as potassium-40 as well as for external radiation such as cosmic rays. What this implies is that the present ICRP recommendation, taken as a basis for public administration by most governments, is that the average exposure of people to artificial sources of radiation should not

exceed roughly half the natural effective dose. What the campaigners are now demanding is a reduction of these limits by a factor of five. At least some of them appear to have calculated that this would be a handy way of pricing civil nuclear power out of business.

The essence of the difficulty is that the factual basis for fixing workable limits is shot through with uncertainty. ICRP's critics say that the urgent question is that of the degree to which small doses of radiation will cause somatic-cell damage leading, among other things, to malignancy. In reality, the most urgent need is to collect the data on which to construct a more accurate relationship between cancer risk and radiation dose at high dose. Only then will it be possible to put a little more flesh on the bones of what is at present known — or, rather, surmised — about the damage done by low doses. For a quarter of a century, it has been conventional to suppose that the relationship between dose and damage is linear — simple proportionality. In other words, it is assumed that there is no threshold of safety. The risk that individuals will contract cancer after exposure to low doses is derived by extrapolating backwards from what is known of the risks of high doses, in the region of 500 mSv and above.

Survivors

Inevitably, the survivors of the explosions at Hiroshima and Nagasaki have been crucial to much of what has been learned in the past thirty years, but sadly, even these careful studies have done very little to pin down the parameters of the relationship between radiation and effect, when the effect is the development of malignancy. One obvious problem is that the numbers involved are very small. Among the group of bomb survivors estimated to have received effective doses between 100 and 200 rem (between 1,000 and 2,000 mSv), for example, the number of cancer deaths between 1950 and 1978 was 222, compared with some 182 expected from natural causes. Plainly the 40 excess cancer deaths provide a point on a graph from which a straight line may be extrapolated back to zero. But it stands out plainly that both the sampling and the systematic errors must be large. ICRP has followed the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR, 1977 report) in supposing that there will be 100 cases of fatal cancer for each million person-rem of exposure to radiation, but the uncertainties of this rule of thumb are vividly illustrated by the roundness of the numbers quoted.

What more can be done? For the past two years, people have been reworking the data for the exposure of people at Hiroshima and Nagasaki with two objectives: to take account of the recognition some years ago that, because of uncertainties about the transmission of fast neutrons through the atmosphere, the doses to which the bomb survivors were exposed may have been overestimated. This will entail reclassifying the 100,000 people (dead as well as alive) in the original study into possibly different categories of effective dose. There is also new data on the energy released by the two weapons. It seems to be accepted that the new data, when complete, will paint a somewhat gloomier picture of the risks of radiation exposure to large acute doses of



radiation than has been current. People in the know think that a revision of the slope of the dose-response curve (in this case a straight line) by a factor of two or so may be the outcome. The first word may come from UNSCEAR, which meets next year. But ICRP's critics are not prepared to wait, and are asking that the backwards extrapolation of this still hypothetical curve should be decreed forthwith.

In reality, in the assessment of the effects of small doses, an element of counting angels on the head of a pin is almost unavoidable. It is possible, for example, to argue that exposing a tissue to an acute dose of radiation will have the effect of killing some of its constituent cells, especially those which suffer the greatest dose, with the result that the incidence of malignancy arising from somatic mutations in the remaining cells will be an underestimate of the damage that would be done by small doses. That argues for a backwards extrapolation of the curve relating effect to dose that rises above the straight line through the origin. The notion that it may require two hits on a DNA molecule to cause a significant somatic mutation may or may not lead to the same consequence. These arguments ignore the effects of DNA repair mechanisms in all living cells, about whose efficiency in different circumstances almost nothing is known.

In the long run, harder evidence may come from careful studies of populations occupationally exposed to radiation, workers in nuclear plants for example, but the need for large numbers (and the long timescale of cancer induction) will make this process slow. A well-run study of those exposed at Chernobyl would also eventually be invaluable. Meanwhile, those seeking further information on the two sides in this disjointed argument could do worse than read *Radiation and Health* (eds Russell Jones, R. & Southwood, R., Wiley, Chichester, 1987). Few will be persuaded that the dose-effect relationship at small doses can be determined unambiguously from the data now available, but many will be stimulated to see how hard evidence might be derived.

Controls

This will not be quickly done. In the nature of the problem of radiation risks at low doses, the numbers must be large and the interpretation for a time uncertain. So much is evident in discussions of the Japanese data now being reclassified. Should the control group be those who happened to live at Hiroshima or Nagasaki, but who were not exposed to significant amounts of radiation? Or should it be the general population of Japan? The first choice offends those who think the evidence already justifies the assumption that radiation effects are greater than simple proportionality would suggest at small doses. The second is plainly wrong, given that both Hiroshima and Nagasaki are (and were) industrial cities.

The determination of the effects of small doses of radiation is a gigantic biological study of the human population. In the long run, the work will be done, and there will be a tangible dose-effect relationship at low doses. It could even prove that the present-day critics are correct in their now-unsubstantiated assertions. (It could just as easily turn out that there is, after all, a threshold, although that seems unlikely on theoretical grounds.) The practical issue is that of whether public administrations should immediately change the rules or wait until more is known. If the radiation exposures of human populations were not still small fractions of the ICRP recommended limits, the case might be different. As things are, the interests of public health would probably be better served by the urgent investigation of a related problem whose importance has come to light only recently — the unexpectedly high concentration of radon in some poorly ventilated modern houses.

Where does all this leave ICRP itself? To respond as its critics ask by promptly tightening the present limits would be as damaging of its reputation and effectiveness as if it caved in to a demand from the nuclear industry that the limits should be

moved in the other direction. It has no choice but to move deliberately. Serious people know that a careful review of the reclassification of the Japanese data must take time, for example. But that is not to say that ICRP should pay no attention to the clamour there has been in the past few months. If it seeks to retain its influence, it had better change its style. For reasons connected with its constitution, but none the less excusable on that account, ICRP is slower than it should be to respond to changing circumstances, and given to behaving as if its recommendations should be regarded as mosaic tablets, to be accepted by all concerned with only the most laconic explanation. It may not have sought the place on the public stage it now occupies, but being there it should learn to act. □

Private public service

Britain is learning again that private monopolies can be as unlikeable as nationalized industries.

LESS than a year ago, the British government seemed to be basking in the knowledge that its schemes for selling public industries to private persons were being so widely imitated elsewhere (in Japan, New Zealand and France, for example) that it could count this imitation as a form of flattery. Now, not unexpectedly, the glow is fading. One of the first and largest industries to be sold to private shareholders, the telephone company called British Telecom, has been through a great barrage of complaint from newspapers complaining that its service is atrocious. The hapless company, saying that it is hampered by the great weight of antiquated electro-mechanical switching gear at its exchanges, appears to agree, promising service of a kind that people know they can reasonably ask for only in 1990 or 1991. The government is in a cleft stick. If it acts tough, and faces British Telecom with the threat of competition on a scale that would force it to be efficient, the value of the privately held shares and of its own holding will no doubt decline sharply, but if it does nothing, the word will get around that selling public companies to private owners is not quite the recipe for prosperity the government has made out.

It will be no bad thing if this little awkwardness gives the government pause. It will no doubt reflect that much of its programme for selling public assets has been successful. Companies such as Amersham International and Rolls-Royce are no doubt in better shape now than when their decisions were monitored (and sometimes made) by civil servants. There is no reason why British Airways, already big and reasonably profitable, should not be allowed in the future to sink or swim in the competition with other carriers by its own efforts. But these are enterprises of a kind which are bound to be made more efficient in an international marketplace in which they are no longer sheltered by government ownership. By contrast, public monopolies, even when faced with limited competition as British Telecom has been (Mercury International is licensed to compete with it, but not to the tune of more than 10 per cent until 1991), have every incentive to make profits, not to provide better service at their shareholders' expense.

Much of this has been clear, in Britain, from the outset of the government's programme. But there is no reason why even the most obdurate of public monopolies should not be sold to the private sector in a way that enhances the incentive to be efficient. Britain owns a number of coal-mines, for example, some of which would be profitable even if the protective barriers of import restrictions were removed. Why not sell them off? More adventurously, why should not the government, which has an ambition to sell the nationalized electricity industry, should begin by selling its power stations singly, not as an integrated network. By British standards, the result would seem an anarchic patchwork of tiny enterprises competing with each other for the business of selling electricity to the distribution network. But is that not what competition is about? □

More money for science to come in Japanese budget

- "Priority areas of research" budget doubles
- Surprise handouts for space and astronauts

Tokyo

JAPAN'S Ministry of Education, Science and Culture has made ambitious requests for its science-related budget in 1988 — the budget is about 20 per cent up over this year — but the Ministry of Finance will almost certainly prune the proposals before submission to the Diet in January.

Requests for grants-in-aid for scientific research are increased by ¥4,000 million (\$27 million). Half of the increase is in the new grant category "priority areas of research" (see *Nature* 323, 284; 1986) which supports projects involving large groups of scientists. Individual grants run for 3–6 years with annual funds of between ¥50 and ¥600 million. From next year, for example, Professor Muto of Tohoku

University will lead a 3-year investigation into the new high-temperature superconductors with a team of 50–60 researchers drawn from institutes throughout Japan.

More than ¥2,200 million (in the category "maintenance of the scientific research system") is requested to reorganize the Astronomical Observatory of Tokyo University into an independent national research institute for joint university use (see *Nature* 323, 574; 1986) — funds for reorganization were requested this fiscal year but were turned down by the Ministry of Finance.

Joint university–industry research now amounts to several thousand million yen per year. A further increase of more than ¥1,000 million is called for in 1988. Part of the increased funds will be used to establish five new "joint research centres", in, for example, the universities of Nagoya and Gumma, to carry out research "directed towards the practical needs of society". Three such centres have been established this year in Kobe, Toyama and Kumamoto universities.

In line with government calls for the internationalization of Japan, the education ministry has allotted ¥373 million (\$2.6 million) under "international academic exchange" for a new postdoctoral fellowship scheme that will be run by the Japan Society for Promotion of Science. In the first year, 50 researchers will be invited from the United States and another 50 from European countries. The ministry also hopes to add 100 new fellowships to its domestic postdoctoral fellowship scheme for Japanese researchers,

1988 Science Budget request for Ministry of Education, Science and Culture

	Yen (thousand million)	Percentage increase or decrease
Grants-in-aid of research	49.1	+ 8.8
Maintenance of scientific research system	2.3	—
Promotion of government/industry research	8.9	+13.0
Research fellowships	1.5	+33.6
Nuclear fusion	7.7	+ 0.4
Accelerator physics (TRISTAN)	17.5	+35.7
Space science	20.6	+74.6
Marine science	5.5	+84.9
Earth science	2.1	+ 1.6
Antarctic research	2.9	0.0
Cancer research	2.2	+19.9
International academic exchange	6.8	+23.3
International student exchange	18.2	+25.5
Japanese language education	0.3	+ 7.5

Superphénix leak located but causing concern

Paris

THE site of a leak of liquid sodium from the fuel rod storage tank of the 1,242 MW Superphénix fast-breeder nuclear reactor near Lyons, France, has been identified. Using an auditory technique first tried at Dounreay in Scotland, technicians draining off the sodium located a hole near the base of the tank on 5 September. Specialists will now be able to determine the cause of the fault. The design of the tank makes repair difficult. It is more likely that the whole tank will have to be replaced, an operation could take three years at a cost of over FF400 million (£40 million; see *Nature* 328, 100; 1987).

At present, Superphénix is not essential

for France's domestic electricity needs. Nevertheless, the French nuclear safety advisory board (SCSIN) will meet on 1 October to decide whether Superphénix may safely be started up again without the defective tank.

Pierre Schmitt, director of Superphénix, hopes that the reactor will come on-stream again while repairs to the tank are carried out. As this is France's prototype, and the world's largest fast-breeder reactor, it is the suspension of on-site training of technicians and the loss of operating experience that are of immediate concern to Electricité de France (EDF), the nationalized utility which runs Superphénix.

Peter Coles

French research budget hints no reassurance

Paris

JACQUES Valade, French research and higher education minister, gave scant assurance to scientists last week when he confirmed that 150 new posts will be created in 1988 in the state research institutions (100 at the Centre National de la Recherche Scientifique, 25 at the Institut National de la Santé et de la Recherche Médicale, 16 at the Institut National de la Recherche Agronomique and 9 at ORSTOM, the scientific research, development and cooperation institute). A minimum of 750 new posts had been recommended by the government-appointed consultative body, the Conseil Supérieur de la Recherche et de la Technologie. Valade also announced that the overall research budget will increase by 8 per cent, compared with 1987. This includes both military and civil spending, however, and the exact allocation of resources within the public sector has still to be published.

Peter Coles

bringing the total number of new 2-year awards to 424 next year. And the number of foreign students under the ministry's scholarship scheme will be increased by 528 to 4,873 with a target of 10,000 by the end of this century.

But the biggest proposed increases are for high-energy physics and marine and space science. The research budget for TRISTAN, the world's largest electron-positron collider at the National Laboratory for High Energy Physics, leaps by more than a third to ¥17,500 million. Full-scale collision experiments began at the end of May and most of the extra funds will be used to increase operating hours from 1,800 hours this year to 2,500 hours in fiscal year 1988.

Funds for marine science are almost doubled, but nearly all the budget is consumed in construction costs for a new research vessel for the Ocean Research Institute of Tokyo University.

The most startling feature of the ministry's budget, however, is the proposal to allocate more than ¥20,000 million to the Institute of Space and Astronautical Science. Funds for the Institute hit a peak of nearly ¥16,000 million in 1983 but have since fallen.

The only new proposal in the space institute's budget is a request for about ¥300 million for another X-ray satellite (ASTRO-D) which will be a follow-up to the recently launched Ginga (ASTRO-C). But it will be surprising if the Ministry of Finance accepts in full this huge increase.

David Swinbanks

No strategy for Space Station

Washington

RECENT criticism of the National Aeronautics and Space Administration (NASA) for its lack of long-term strategy continued this week with the publication of a report, by a committee of the National Research Council (NRC), on the Space Station programme. Although the report commends NASA's plan for the initial phase of the project as an appropriate first step, it finds the proposed second phase poorly thought out, and that deployment using the space shuttle, in its presently anticipated design, has serious limitations.

Phase I of the space station is a long boom, with laboratory and habitation modules in the centre and solar arrays at each end. It will provide a 'microgravity' laboratory for materials science and biology experiments, and the NRC committee pronounced it a reasonable way of achieving its intended aims. But Phase II, which adds a servicing facility and side booms to accommodate scientific instruments, does not "reflect the right priorities for space station evolution". Its equatorial orbit makes the space station a poor platform for Earth observation, and for astronomy free-flying spacecraft are preferable. The servicing facility will be useful for a new generation of large orbiting structures, but insufficient for making the space station into a stepping-stone for manned lunar or planetary missions.

A more immediate worry is simply getting the space station aloft. The shuttle will, according to NASA, put the first payloads into an orbit about 150 miles high, where atmospheric drag will bring an unpowered vehicle down in only 20 days. Any failure of the reboost system during assembly would be catastrophic. Possible remedies are reduction in weight of the first payloads or, in the longer term, use of a more powerful shuttle or a new heavy-lift vehicle.

The report finds NASA's accounting of the test and safety procedures inadequate, and that extra costs of up to \$4,700 million could arise in the provision of prototype and backup hardware.

The committee stresses that recent experience has taught that space programmes cannot be done 'on the cheap'; testing should not be skimped, and other space science activities should not be straitjacketed to conform to the limitations of the space station programme. Throughout the report is an underlying theme that NASA lacks a coherent strategy, and the committee recommends that before going too far with this, or any other big project, there should be a clarification of the long-term goals of the US space programme. **David Lindley**

UK remote sensing programme latest victim of cutbacks

London

AMID rumours that the British government may be preparing to soften its stance on the question of space funding, scientists from the Natural Environment Research Council (NERC) in London last week to publicise the potential applications of the emerging technology of remote sensing — the use of satellites to observe the Earth from space. Such possible applications, says NERC, range from the prediction of drought and the control of locusts to vegetation and urbanization monitoring.

For the British National Space Centre (BNSC), which coordinates civil space research in Britain, remote sensing is a priority. Of BNSC's £80 million contribution to the European Space Agency, ESA, £22 million goes on remote sensing. A further £14 million is spent domestically, mostly on the Royal Aircraft Establishment (RAE) at Farnborough, since 1980 the home of the national remote sensing centre.

In mid-1990, ESA is due to launch its first remote-sensing satellite, ERS-1, which will have a projected lifespan of three years, after which a successor, ERS-2, will replace it. To use the information supplied by the ERS satellites, an ERS data centre is to be set up at the RAE. When the government announced in July that there would be no new money for the national space plan (see *Nature* 328, 467; 1987), a decision that prompted the resignation of BNSC's director-general, Roy Gibson (see *Nature* 328, 565; 1987), one aspect that was largely overlooked during the ensuing debate was the future of the ERS data centre.

The research teams charged with identifying the sorts of systems needed fully to exploit the ERS data originally calculated that £17 million of government funds (ESA is to supply £3 million) would be required before the launch of ERS-1. It seems that BNSC can now afford to provide only £4 million annually for the next five years, allowing a total of only £8 million for the pre-launch facilities. This reduced rate of funding has forced RAE to reconsider its priorities. Many data products with potential commercial applications will have to be shelved.

Part of the centre will comprise a processing and archiving facility, primarily for ESA's use. The Department of Trade and Industry, BNSC's parent body, has in the past made it clear that it would like to see such a facility made available to UK users. Without the extra money, priority will be given to ESA.

Because the money for the data centre will not be forthcoming until next April,

the teams of scientists responsible for laying the groundwork, employed by RAE on a contract basis, are in danger of breaking up. The product support team's present contract expires in October. If it cannot be renewed until April, it is likely that several of the team members will be redeployed, possibly irretrievably.

The space community in Britain has not yet given up hope. Last month the government gave BNSC an extra £4 million, having a month earlier said that no new money was available. There is a feeling that the government may further relent after next month's meeting of the newly formed Advisory Council for Science and Technology. **Simon Hadlington**

Deep-ocean disposal plans jettisoned

London

MEMBER governments of the Paris-based Nuclear Energy Agency (NEA) have discontinued a five-year research programme aimed at proving the feasibility of disposing of high-level radioactive wastes in deep-ocean sediments and even crystal rocks. The governments, all members of the Organization for Economic Cooperation and Development (OECD), of which NEA is an off-shoot, say that deep-ocean disposal is too far in the future for further research to be justified.

Jean-Pierre Olivier, head of waste disposal at OECD, says that much of the work of NEA's Seabed Working Group is "very exciting", but is now pessimistic about continued research because of "difficulties on a political level". Among the schemes investigated are sinking projectiles laden with waste into ocean sediments and the deeper burial of wastes in boreholes.

Reactions are typified by those of the British government, where high-level wastes are soon to be vitrified and stored at land sites for 50 years. The chief inspector of radiation pollution, Dr Frank Feates, at the Department of the Environment, asks whether more research is needed "when the reports will sit on the shelf for 50 years".

The ending of the programme has nevertheless catalysed a financial crisis at British oceanographic laboratories. The Institute of Oceanographic Sciences, which earned £1.6 million a year from the deep-ocean programme, is unclear whether it should have anticipated the ending of this contract last March (see *Nature* 326, 96; 1987), but the Department of the Environment believes it more likely that the news did not get through clearly enough when budgets were being drawn up.

Kathy Johnston

Two wins for environmentalists in Australian wilderness battle

Sydney

In the past week Australia's federal Labor government has won two battles in its campaign to conserve the cream of the continent's wilderness areas in conservatively governed states. In the Northern Territory, a court gave the federal government permission to proceed with the nomination of the 6,929 km² of Kakadu National Park Stage 2 for inscription on the UNESCO World Heritage list. And in Tasmania, a court upheld the federal government's right to prevent logging or road construction in an area also being considered for World Heritage listing over the state government's objection.

In the Kakadu National Park, a mining company which felt that its interests had not been properly taken into account had sought to block the World Heritage listing. The company holds exploration leases in the area thought to contain uranium as well as gold, platinum and basic metal deposits. World Heritage status would forbid both mining and exploration (*Nature* 326,434; 1987).

The Tasmanian area is virgin cold-temperature rain forest. It is the habitat of the world's tallest flowering plant, the Swamp gum, which can grow up to 99 m high.

The next target for listing is likely to be 9,300 km² of tropical rainforest along the north coast of Queensland near the town of Cairns. According to the federal minister for the environment, Senator Graham Richardson, the north Queensland rainforest contains the highest concentration of primitive flowering plants in the world and holds important clues to the origin, evolution and migration of flowering plants. Australian conservationists maintain that even though the size of the north Queensland rainforest is small by world standards, it is the only tropical rainforest inside a developed country.

But the right-wing premier of Queensland, Sir Johannes Bjelke-Petersen, sees the World Heritage listing as a takeover attempt by the federal government and has vowed to do everything within his government's power to prevent it.

Charles Morgan

First steps in ozone protection agreed

London

A historic agreement to protect the Earth's ozone layer is due to be signed this week in Montreal, the culmination of months of negotiations over the specific timing and severity of regulations to limit the man-made chemicals chlorofluorocarbons (CFCs) and halons.

During last week's tough negotiating session on several contentious issues involved in the protocol, it looked at times as if the protocol might break down completely, but always it was steered back on track. However, even as ministers and government representatives from some 35 countries arrived for the final session, some outstanding issues remained.

The final document looked set to provide a 50 per cent cutback in the consumption of CFCs 11, 12, 113, 114 and 115 by the beginning of 1999, following a freeze in 1990 and a 20 per cent cut-back in 1994. Halons 1211 and 1301, widely used as fire extinguishants, will be frozen at 1986 levels by 1993.

Research results expected in a matter of weeks on both Antarctic and global ozone depletion should help shed some light on the question of whether the new regulations go far enough, and the United Nations Environment Programme director Dr Mostafa Tolba has reserved the right to call an emergency meeting if necessary. No one was willing to postpone the protocol signing, however, for fear of upsetting its

delicate momentum.

Insistence by the United States, Canada and Nordic countries on the need for a 95 per cent reduction in CFCs is thought to be largely responsible for the shape of the final agreement with its 50 per cent cutback; considering the position of some countries, including Britain, at the start of negotiations last December, a mere freeze on CFC production was another possible outcome.

Many hours of intense discussions in Montreal over the issue of CFC use by developing countries resulted in a compromise allowing a ten-year grace period during which these low-consuming countries would be allowed to reach a ceiling of 0.3 kg per person of CFCs (compared with an estimated 0.8 kg in developed countries).

In a compromise between the United States and the European Community countries, CFC limits will apply to consumption, rather than production, which would still allow exports of the chemicals.

The actual numbers included in the protocol specifying percentages are recognized as politically rather than scientifically based but the agreement is being hailed as an important first step in ozone protection. It remains to be seen whether results from the latest Antarctic expedition and studies by the US National Aeronautics and Space Administration will demonstrate a need for further restrictions.

Kathy Johnston

Swedish view of US science policy tangles

Washington

Dr Frank Press, president of the US National Academy of Sciences, last week took on the daunting task of describing US science policy to the Swedish Prime Minister, Ingvar Carlsson, during his state visit to Washington. After describing scientific enterprise in the United States as "decentralized, some would say disorganized, and yet successful", Press led a lively if eclectic 100-minute discussion staged at Carlsson's request.

Not wanting to shoulder the burden alone, Press asked James Wyngaarden, director of the National Institutes of Health, and Erich Bloch, director of the National Science Foundation, to explain the role of civilian agencies in supporting science. Representatives of defence research



Press (left) and Carlsson — future possibilities.

funding agencies were notably absent.

Sam Thier, president of the Institute of Medicine, and Robert White, president of the National Institute of Engineering, described how independent scientific organizations could advise government. Carlsson asked if their advice was heeded, but White dodged the question, explaining that government agencies had usually asked for the advice in the first place.

One immediate issue touched on and probed more deeply at a later meeting without the prime minister was data transfer. Sweden will house the secretariat for the International Geosphere/Biosphere Programme (IGBP), making it the focal point for a huge volume of data. Earth observation satellites from several countries will produce a steady deluge of telemetry that will need to be collected and distributed to make IGBP a success.

Carlsson told his hosts that he met regularly with members of the Swedish scientific community, often once a month. Scientists present possibilities, said Carlsson, whereas most of a prime minister's job consists of coping with problems.

Joseph Palca

US AIDS education programme on-again, off-again

Washington

PLANS for a coordinated, multi-agency education campaign next month aimed at slowing the spread of AIDS (acquired immune deficiency syndrome) are in disarray. Public service announcements await approval, publicity for the campaign has been muted and a national mailing of an information brochure has been scrapped. But the Centers for Disease Control (CDC) are pushing for an "AIDS Prevention Month", despite conflicting signals on how to proceed from the highest levels of government.

CDC director James Mason described some of his agencies' plans to the presidential AIDS panel that last week held its first meeting (see right). The CDC campaign's theme is "America responds to AIDS", and will involve government agencies working in partnership with state, local and private groups already coping with the AIDS pandemic. CDC is also sponsoring forums to discuss how to make these efforts more successful. Five cities will receive a heightened level of attention, including underwriting for television programmes, information desks in pharmacies and more intensive public relations campaigns.

But as initially conceived, CDC's activities were only one part of a government-wide education effort. Paula Van Ness, director of the national AIDS information programme at CDC, says an educational brochure was to be mailed to every US home as part of the month-long education blitz. But a decision to go ahead with the mailing failed to materialize, despite the fact that Congress had appropriated some \$20 million to cover the cost of the campaign. That money must be spent before 30 September, the end of the current fiscal year, or Congress must give its permission to use it for something else. The administration is currently debating an alternative method for distributing the brochure, targeting groups with a special need for education. The issue of whether to mail the brochure was on the agenda of the presidential AIDS panel, but was removed at the last minute.

Robert Sweet of the Domestic Policy Council says plans for a national mailing have been on hold for some time, pending the outcome of a national survey being conducted by the National Center for Health Statistics (NCHS) to assess the public's knowledge and attitudes toward AIDS. Beginning in August, a series of questions was added to the National Health Interview Study — a 30-year-old continuing survey on various health topics — to determine what Americans already know about AIDS. Each week, some

1,000 households are surveyed, with an 88 per cent response rate so far. The first published data are expected in November, but Owen Thornberry of the NCHS says two-thirds of respondents have indicated they would be willing to donate blood for the national random AIDS antibody test being planned by CDC. The NCHS survey will be used to determine the effectiveness of government education campaigns.

The October AIDS Prevention Month is one of the first major campaigns whose effects are expected to show up in the survey. Although plans are still being formulated, CDC is working on 31 events — one for each day of the month — with different organizations around the country. But in every case, the events have been planned with a host group that can continue the event if CDC is instructed to call off their plans. Chuck Klein, a spokesman for the Department of Health and Human Services (HHS), says that if Congress declares October national AIDS awareness month, HHS will join in with enthusiasm. HHS Secretary Otis Bowen has recorded a series of public service announcements about AIDS, and the department is negotiating with the National Association of Broadcasters to present a multi-part radio programme on stations throughout the country.

Another problem frustrating government-sponsored education plans is the inability to purchase advertising time during prime viewing hours. The government has contracted with one of the largest US advertising agencies, Ogilvy and Mather, to help develop an advertising campaign. Klein says government agencies are prohibited from purchasing television time with only a few exceptions — the annual \$155-million advertisement campaign for armed forces recruits being the most notable. But the government could find ways around this restriction — soliciting a non-governmental co-sponsor — and Congress has been willing to amend laws to allow AIDS education plans to proceed.

The "National AIDS Awareness Test" broadcast on television stations nationwide on Tuesday evening this week was not encumbered by the laws restricting government advertising. It was sponsored by the Metropolitan Life insurance company. The hour-long show had a question-and-answer format, with answer forms printed in programming guides so viewers could record and check their scores. The show featured Surgeon General C. Everett Koop, and other prominent public health and medical officials, so the government's educational agenda was well represented.

Joseph Palca & Carol Ezzell

AIDS clashes in presidential commission

Washington

GAY activists, government agency officials, special interest groups, and health officials clashed last week at the first meeting of the US Presidential Commission on the HIV Epidemic, named after the virus that causes AIDS (acquired immune deficiency syndrome). The panel heard how the Public Health Service and other federal agencies are combating the AIDS pandemic, but they also listened to testimony suggesting that the government is not doing enough to counter the crisis, and that federal efforts are disjointed and suffer from credibility problems.

The AIDS commission has been under fire since its formation by presidential order in June (see *Nature* 328, 366 and 372; 1987) for lack of scientific expertise, and for the social and moral views of some of its members. Speaking before the panel, Otis Bowen, Secretary of Health and Human Services, defended the commission, adding that "constant criticism that this government is not doing enough is unfair and unwarranted".

Surgeon General C. Everett Koop told the commission of his concern about discrimination against patients with AIDS. Koop had strong words for doctors and other health professionals who refuse to treat AIDS patients, saying their attitudes threaten the basic assumption in medicine that health care should be available to anyone who needs it.

A parade of agency chiefs and department heads described to the panel the spectrum of federal activities to prevent the spread of HIV while preserving the rights of those already infected. At the end of the presentations, panel member William Walsh, founder of Project HOPE, pointed out that many federal programmes seemed to overlap, particularly in the area of educational campaigns. When he asked Ralph Bledsoe of the White House Domestic Policy Council why there seemed to be a lack of coordination between the efforts of different agencies, Bledsoe replied, "welcome to the federal government".

The commission has much to do before it prepares its preliminary report in December, and its final recommendations to the president by next September. Commission chairman Eugene Mayberry, head of the Mayo Foundation, urged the press and the public not to "make too much of controversial issues until we've had a chance to prepare". The commission is scheduled to meet in New York, San Francisco and Nashville during the next year.

Carol Ezzell

UK edges slowly towards first trials of an AIDS vaccine

London

Six months after initiating a directed programme of research on AIDS (acquired immune deficiency syndrome), the UK Medical Research Council (MRC) is confident that its first vaccine trials will begin within a year. Research into new drugs will take longer to bear fruit.

The £2.5 million earmarked for the first year of the programme has now been committed and two major grants, each of £1 million over five years, have been awarded. One grant is for Professor Robin Weiss of the Institute of Cancer Research to study the biology of human immunodeficiency virus (HIV) and to carry out HIV neutralization tests as a service. The other grant is for Professor William Jarrett's work at Glasgow University on viruses and vaccines.

Jarrett favours subunit vaccines, in which the essential component is a purified viral protein, rather than vaccines based on vaccinia viruses carrying an HIV gene. Monkey trials of such vaccines have all failed, he says, and his own work on the use of adenovirus in place of vaccinia virus has encountered a serious obstacle. Moreover, he would now be concerned by the prospect of any vaccine that led to the continued synthesis of viral envelope protein in view of its possible toxicity.

Jarrett and his brother Oswald have developed a highly effective vaccine against feline leukaemia virus, which is a retrovirus like HIV, and have already tested a similar HIV vaccine in monkeys. In the vaccines a protein from the envelope that surrounds the virus is incorporated into an 'iscom' (immune stimulating complex).

Iscom technology, developed by Bror Morein of the Swedish University of Agricultural Sciences in Uppsala, attaches the viral protein by hydrophobic interaction to a particle-like matrix of the glycoside Quil A, which is extracted from the bark of Amazonian oak trees. Iscom preparations induce a greater immune response than most alternative vaccine preparations. (The latest potential rival is based on virus-like particles produced by yeast — see *Nature* 329, 68; 1987.)

Jarrett and Sir James Gowans, who retires as head of the Medical Research Council this month, are both confident that the way will be cleared to carry out initial human tests of an HIV-iscom vaccine within the next year. Tests on monkeys, carried out in conjunction with the US National Cancer Institute, have shown that HIV-iscom preparations are both safe and good immunogens. The favoured protein for the first human tests is the gp120 envelope protein of HIV, the

major subunit of the gp160 protein that will be used in the first US vaccine test (*Nature* 328, 747; 1987).

The main hurdles to an early UK trial are now bureaucratic. General guidelines, an approved protocol, ethical issues and the choice of volunteers have all to be decided. Both Jarrett and Dr Geoffrey Schild, who is to succeed Gowans as the director of the MRC's AIDS programme, emphasize that they are determined to avoid vaccine trials that are carried out by "enthusiasts" on a small number of people and using unstandardized protocols. One would be little the wiser from

such trials, said Jarrett. The task of standardization will fall to Schild, wearing his other hat as director of the National Institute for Biological Standards and Control.

On the antiviral front, the MRC programme has been less immediately productive. Its director, Dr Max Perutz, says that one aim is to arouse the interest of academic chemists in new approaches to the synthesis of inhibitors of HIV's reverse transcriptase enzyme. As befits his own experience, Perutz is also keen to support attempts to determine the crystal structure of HIV proteins so as to allow more rational drug design. Projects to crystallize HIV reverse transcriptase, proteinase, envelope and tatIII proteins are being supported in Cambridge, London and York.

Peter Newmark

Safety issues raised by worker's infection with AIDS virus

Washington

A LABORATORY worker who became infected with human immunodeficiency virus (HIV), the virus causing AIDS (acquired immune deficiency syndrome), may not have followed accepted safety procedures while handling highly concentrated samples of the virus (see *Nature* 329, 92; 1987). That conclusion follows a telephone interview between the infected worker — whose identity is being withheld to protect confidentiality agreements — and the leader of a safety team investigating the circumstances surrounding the infection first detected 18 months ago.

Emmett Barkley, former director of the division of laboratory safety at the National Institutes of Health (NIH) and leader of the safety team says the worker could not identify an overt event such as a broken vessel or needle stick that could have resulted in transmission of the virus. But Barkley says his conversation revealed several possible "inapparent exposures". The worker, whose job included concentrating the virus in a centrifuge using a density gradient, reported that the seals around the centrifuge rotor had occasionally leaked. The worker handled the virus at times when skin lesions, cuts and abrasions were present, but always wore gloves while handling the virus. But Barkley says gloves may tear or have pin holes without the wearer being aware of it.

The infected worker did not work on campus at NIH, but instead was employed by an NIH contractor. Peter Fischinger, deputy director of the National Cancer Institute, says he believes that NIH safety procedures are adequate, but agrees that the incident raises questions about whether workers are adequately trained in procedures. He also says some modifica-

tions of procedures for handling the virus may be made. NIH officials are convinced that the infection did not result from inhalation of an aerosol.

Barkley says NIH rules do not require that HIV be handled in P3 facilities. (These high-level containment facilities are now officially known as BL3 facilities, for biosafety level 3.) Instead, they allow the virus to be handled in BL2 facilities using BL3 techniques such as wearing gloves and working in a laminar flow hood.

Some at NIH have criticized the institute's handling of the event. Fischinger says the Centers for Disease Control (CDC) in Atlanta were notified of the occupational exposure after the virus was isolated in the infected individual and shown by restriction mapping to be essentially identical to the one being handled in the laboratory. But that determination was made only last month, more than a year after the individual had shown evidence of antibodies to HIV.

NIH director James Wyngaarden admits the delay in informing CDC may have been a mistake, but says it was caused in part by the difficulty in isolating the virus from the infected worker.

Wyngaarden says nine unsuccessful attempts to find the virus in the worker's blood made researchers question whether the antibody testing might not have been falsely positive for the presence of virus. Only after investigators tried a test developed by Robert Gallo was the virus successfully grown.

NIH safety inspectors are now in the process of visiting all facilities around the country where concentrated quantities of HIV are produced. This includes companies making the HIV-antibody test kits.

Joseph Palca

South African government threatens university subsidies

Cape Town

SOUTH AFRICA'S universities are anxiously waiting to see whether the government seriously intends to carry out its threat to cut the subsidies on which they all depend. The first sign of impending trouble came on 5 August, when the Minister of National Education, F.W. de Klerk, summoned all university principals to tell them of the government's dissatisfaction with political activity on campuses.

Since then, the relevant ministers of education (for the various racial groups) have sent letters to the chairmen of university councils demanding that universities themselves act to prevent disruption of academic activities. They are also required to prevent staff and students from using university facilities to promote civil disobedience of any kind, support any boycott or work stay-away, organize any illegal gathering or promote the image of any illegal or 'affected' organization.

This last provision is clearly aimed at restricting the National Union of South African Students (NUSAS), which was declared an 'affected' organization in the mid-1970s, thereby preventing it from receiving funds from overseas. The student representative councils of the Universities of the Witwatersrand, Cape Town, Rhodes and Natal are affiliated to the union, which would be unable to function if it were denied its campus base.

The government is demanding that the universities discipline any student or staff member who engages in prohibited activities or conducts him or herself in a seditious manner within 2 km of the perimeter of the campus. Furthermore, it requires that the universities inform the relevant minister of education of any incident of unrest and any occurrence of such activities, and of the disciplinary steps taken. If the universities fail to meet the minister's requirements, he threatens to enforce a subsidy cut. All South Africa's universities receive a subsidy from the state, which amounts to at least 75 per cent of the revenue of each university.

The minister's threats are clearly directed primarily at the Universities of Cape Town (UCT) and the Witwatersrand, both of which have a substantial minority (about 20 per cent) of black students. Two incidents took place on the UCT campus last month. In the first, a group of students at one of the residences disrupted a speech by the former South African ambassador in London, Dr Dennis Worrall. (Subsequently, one student was suspended until the end of the year and several others fined by the university court.) The following day, Mr Tom Linda, ex-mayor of Ibhayi, a Port Elizabeth township, was

chased off the campus by students after having arrived to address a meeting of the so-called Moderate Students Movement. The meeting was held in defiance of a ban imposed on it by the UCT principal and vice-chancellor, Dr Stuart Saunders, on the grounds that the procedures laid down by the university for inviting controversial speakers to campus were not followed.

University councils had until the end of August to respond to the government's proposals. Although these responses have not been made public, it is widely understood that the five open universities have rejected the proposals as being completely unworkable. There has apparently also been significant opposition to the proposals from academics at Stellenbosch, the country's leading Afrikaans-medium

university. There is some speculation in academic circles that de Klerk may back down on his threat to cut subsidies. The possibility also exists that the minister is threatening to exercise powers that are *ultra vires*, and could be tested in court.

While the government is claiming to be interfering in university autonomy in the interests of avoiding unrest, it seems to be employing people with the aim of fuelling precisely such incidents. Saunders met Minister of Law and Order Adriaan Vlok last week to discuss the recent disclosure by a second-year social science student, Daniel Pretorius, that he was a paid security-police spy operating at UCT. This was admitted by Vlok. Affidavits have subsequently been collected by the university alleging that he has acted as an *agent provocateur* campus more than once, including an incident at a demonstration earlier this year in which he encouraged other students to throw stones at a police vehicle.

Michael Cherry

East and West Germany claim to open border for scientists

Bonn

EXPORT restrictions notwithstanding, East and West Germany signed an agreement in Bonn on 10 September for scientific exchange in all fields.

Until now, scientists from West Germany have rarely been allowed to visit their East German colleagues, let alone to collaborate with them. The new agreement also permits East German scientists to travel out of the Soviet bloc.

The new agreement, signed after an exhausting 14 years of negotiations, calls for a joint commission to oversee cooperation. The West German Research and Technology Ministry (BMFT) announced that 27 projects have already been agreed. The commission will sit starting on 1 November to settle the details of these projects and to decide on other areas for cooperation. The 27 projects include research on AIDS, nuclear and non-nuclear energy, and biotechnology.

Both sides face major obstacles in implementing the agreement, signed last week during the historic first visit to West Germany of the East German leader, Erich Honecker. West German industry has run into trouble before with Western restrictions regulating what may be exported to countries in the Soviet bloc.

It is hard to imagine that collaboration in a field such as computer-integrated manufacturing, included in the initial list of 27, could possibly go smoothly. But the BMFT played down the possibility that its Western partners might balk at the agreement. BMFT official Karl Christoph Blaesing pointed out how few of the first 27 projects involve high technology. East

Germany, the largest computer exporter in the Soviet bloc, may have a few secrets of its own. But East German Science and Technology Minister Herbert Weiz said there are "no limitations" on what can be discussed with West German scientists. "No one gives orders to our sovereign state" about what it can buy or sell, he said.

Some are sceptical of just how open East Germany will be. Nobel physics laureate Klaus von Klitzing, says "I will try to invite a particular scientist and see if he, not his boss or somebody else, really comes."

Steven Dickman

Vera Rich adds — The East German view of last week's agreement is that it is but part of a continuing process of detente. East German sources point out that many proposed areas of cooperation such as energy, environment and health, had been scheduled for discussion at the 1980 Hamburg Science Forum.

East Germany is now more conscious of environmental issues, and has been cooperating with West Germany in several projects, even agreeing that West German organizations with headquarters in West Berlin should be allowed to take part despite their constitutional non-existence in East German eyes.

The high-technology parts of the agreement are likely to be the most interesting to East Germany, which is opposed to western restrictions of collaboration in these fields but was nevertheless quick to insist last week that it has the intellectual capacity to render them nugatory. But the prize for East Germany would be reduced dependence on Comecon, the Warsaw Pact's economic organization. □

Soviet academy condemned for "lack of democracy"

London

THE Soviet Academy of Sciences, which for decades has been praised abroad as the one remaining institution in the Soviet Union to preserve multi-candidate elections by secret ballot is now being attacked for its "lack of democracy". Comments by scientists in *Literaturnaya Gazeta* last week call the academy a "stronghold of obsolete thinking" and a "fortress against the democratization of scientific activity". The feature appeared on the same day that academy officials met to discuss "restructuring" and democratization — without, apparently, touching on the main cause of discontent.

The academy is not merely a self-electing body of a few hundred eminent scientists; it is also one of the main employers of scientists in the Soviet Union, with a vast network of research institutions and a scientific work-force of tens of thousands. These scientist employees now wish to use the new 'democratization' drive to elect their own directors and heads of department, section or laboratory.

But the president of the academy, Dr Gurii Marchuk told *Izvestiya* last March that if the right of electing people to these posts is given to the scientific work-force, "it could lead to negative consequences". The collective may "appraise" the nominations for director, he said, but it is the scientific council of the institute that must decide.

Marchuk's remarks drew an angry reply from the scientific community. In an 'open letter' published in *Literaturnaya Gazeta*, 22 eminent researchers pointed out that, under the current provisional statutes of the academy, the director "determines" the membership of the scientific council. As this council then elects the heads of department, such posts are usually filled by nominees of the directors.

The presidium of the academy has issued a formal reply saying that democratic traditions have long existed in the academy and will continue to be improved. A commission has been set up and will this month submit a programme to be discussed by the presidium, then handed down to the institutes' collectives for "broad discussion and the subsequent adoption of a definite decision in accordance with the academy's statutes". The institute director will continue to be elected by secret ballot at a general meeting of the relevant branch; the scientific council of each institute, subject to confirmation by the relevant branch and heads of department, will be elected by the work collectives; and at all levels, candidates for election must submit a proposal for the work to be carried out under their direction.

This, however, is not enough, according to the scientific editors of *Literaturnaya Gazeta*. They point out that the new Law on State Enterprises specifies that the enterprise leader is to be elected by the work collective. To deprive scientists of a right now enjoyed by workers and collective farmers is, the editors conclude, "simply offensive". *Literaturnaya Gazeta* has received numerous complaints from scientists about anti-democratic practices within the academy, a resumé of which has now appeared. Cases have been reported of "direct persecution" and "arbitrary behaviour" on the part of institute directors. A letter from A. Fedorov, a Moscow biologist notes that "institute directors are often chosen not for profes-

I have been authorised to tell you that the academy is perfectly democratic.



sional qualifications ... but on the basis of favouritism", and that once in an administrative post, they automatically move up the academic ladder until they become full members of the academy, with honorary titles.

Recently, when vacancies have arisen for institute directors, in virtually every case institute staff have demanded the right to elect them themselves. There have already been some dramatic battles — and in at least one case the staff have managed to reject the official nominee and have their own choice appointed.

The basic difficulty appears to be the contrast between the academy's original role as a small, self-electing elite and its current function as the principal coordinator of scientific activity for the entire Soviet Union. Although this new role is regularly stressed in political exhortations, it has never been reflected in the academy statutes, which even in their latest revision state that "the Academy consists of Full Members and Corresponding Members", ignoring the thousands of scientists employed in academy institutes. And these scientists are unwilling to be ignored any longer.

Vera Rich

Production of recombinant insulin begins

Munich

AFTER a two-year delay, the West German *Land* (federal government) of Hesse on 9 September granted permission for Hoechst AG to produce synthetic human insulin from genetically manipulated bacteria. This is the first time a factory dependent on recombinant bacteria has been approved in West Germany.

Synthetic insulin produced from bacteria causes fewer health problems than the pig insulin commonly used by diabetics. Using bacteria containing a human gene to produce insulin is not a new process; British and US companies have been doing it with little fanfare since 1982. But West Germany's Green party has delayed government approval for part of the Hoechst production facility since 1985.

The insulin plant, thought to have cost Hoechst \$60 million, "included more safety provisions than we required", said Gerd Hobom of the Zentrale Kommission für Biologische Sicherheit (ZKBS), the government organization responsible for assessing the safety of genetic engineering procedures. The bacteria are held in a closed vessel with a built-in system to warn of leaks and are killed before they are removed from the fermenter for processing. The insulin-containing brew is heated once again to denature remaining DNA.

West Germany has no law regulating use of genetically altered organisms. Until now, researchers and companies have voluntarily called in the parent of the ZKBS, the Bundesgesundheitsamt (Federal Health Office or BGA) for a safety rating. The pharmaceutical industry organization has accepted the BGA's authority in this matter; all researchers funded by the federal or *Länder* governments are also required to comply.

But the situation will soon change. The Bundestag (federal parliament) has called on the Health Ministry to draft a law to regulate genetic engineering. Despite the opposition of some scientists who think self-regulation is enough, some version of the bill will probably pass in the coming year.

One other West German company, Boehringer Ingelheim, has used recombinant techniques to create a pharmaceutical product, in this case the heart drug TPA (tissue plasminogen activator; see *Nature* 327, 450; 1987). TPA is produced with mammalian cells in culture rather than with bacteria. Boehringer has already received unofficial approval for producing TPA at its plant in Biberach (which is not in Hesse); written approval is expected any day now. Steven Dickman

Economy in space

SIR—Your leading article "Economy in space" (*Nature* 328, 562; 1987) calls for comment. No doubt what is and is not worthwhile in space deserves scientific evaluation, especially as both this and high-energy physics seem to be coming to the end of their tether. But this is not what the column offers. Instead we have an amazing collection of *non sequiturs*.

First, if Blue Streak was abandoned because it "would make Britain more vulnerable", why do we now have US missiles on our soil over which we have no control? Further, if Britain has "neglected its technical culture over four decades", how is the "present government attempting to redress the balance", by dismantling higher education?

The fact is that Britain has lost its pre-eminence in technology through colossal mismanagement and waste of human resources. Some of us who worked on the scientific war effort and moved to industry have seen this at first hand, finding it impossible to do anything. Later, having emigrated to the other side of the Atlantic, one saw this even more clearly.

Bashing Labour, and now academics, is not going to solve the problem, nor is monetarism, which gives only the appearance (fragile indeed) but not the substance of prosperity. In fact the fear of inflation, referred to in the column, was due precisely to the chancellor having allotted the budget surplus to tax rebate rather than to education and research. Apart from their shortsightedness, these policies have simply given a new twist to the polarization (us and them) that has been the main source of our problems in the first place. True we do have a cultural crisis, but it is not what the column indicates.

M. C. GOODALL

3 Robson Terrace, Shincliffe Village,
Durham DH1 2NL, UK

SIR—Your leading article "Economy in space" (*Nature* 328, 562; 1987) shows a surprising superficiality.

It is your prescription that gives cause for concern. You are right to identify the dilemma that exists for this or any advanced nation regarding its role in space or in any scientific or development activity. But Britain has an advanced economy together with a well-established expertise in space and while much of your previously expressed opinion has concentrated on how to maintain this situation, you are now counselling a national descent to underdeveloped status.

The demise of Blue Streak as a purely British project was indeed inevitable. The inability of Britain to secure a worthwhile involvement in the later successful adoption of that technology in the Ariane programme was not. Britain and other European countries, including Belgium, can through the European Space Agency (ESA) continue to share in a worthwhile European role in space at a cost that each can afford. Britain's relatively minor part in the highly successful Ariane programme followed from the confused and fragmented space policy of successive governments. The formation of the British National Space Centre and the arrival of a space professional such as Roy Gibson to head it were signs that at last a UK government was determined on an effective and coordinated space policy. As the main ESA programme is an optional one, the establishment and operation of such a policy can provide an excellent defence against involvement in extravagant or unrealistic projects.

A range of communication satellites for civil and defence use, a variety of microwave and other sensors together with a demonstrated and world leading position in ground control systems, navigation systems and software are distinctive products of high added value that form an

important part of an industrial sector with a turnover of more than £2,000 million. The United Kingdom's leading role in constructing the Giotto spacecraft which passed within 500 km of the nucleus of Halley's comet surely demonstrates a substantial level of technical competence. Roy Gibson's message to the government was that the industrial returns from our present modest ESA programme could be most effectively enhanced by increasing the complementary UK activity by the equivalent of an additional £2 per UK citizen each year.

This request is far from a demand for a daring role in an ambitious space programme but stems, for example, from a recognition that an involvement in space, in addition to stimulating interest in challenging technological projects, may ultimately be essential to our survival as a species, given our need to monitor, understand, protect and control the climate and the renewable resources of the planet. Increased support of space by industry would be very welcome but is unlikely to be forthcoming for activities which involve a national rather than a corporate return on a timescale of ten or more years.

UK technical people may indeed be undervalued. To "put space on the back burner for a few decades" will do little to increase the perception of the value and importance of technology among ministers, the general public or even the engineers and scientists themselves. Once specialist skills and expertise are lost, particularly if established teams are broken up, they are impossible to resurrect. If we wait two or three decades it will be much too late to repeat Gibson's question.

J. LEONARD CULHANE

Mullard Space Science Laboratory,
University College London,
Holmbury St Mary,
Dorking,
Surrey RH5 6NT, UK

Databases

SIR—C.O. Pabo's letter¹ on molecular biology databases is pertinent; his proposals have been considered at one time or another by those involved in maintaining sequence databases. His "higher order databases" (consisting of structural motifs, similarities to other sequences and the like) are feasible now.

Yet there is no consensus among molecular biologists and the agencies that fund such work that such an advanced database is needed; nor is there support for such databases among the organizations responsible for funding the national sequence databases. My understanding is that the financial support for GenBank² involves the National Institute of General Medical Sciences; the National Library of Medicine; the National Cancer Institute; the National Institute of Allergy and

Infectious Diseases; the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases; the Division of Research Resources at the National Institutes of Health; the National Science Foundation; the Department of Energy; the Department of Agriculture; and the Department of Defense. With 10 organizations supporting the effort, we are still left with a database whose utility is limited because essential information (strain or species-specific variation in sequence, for example) is not accessible or, frequently, not even included.

The problem does not lie with the organizations doing the data collecting; they are understaffed as a result of the meagre commitment of the foregoing organizations relative to the size of the problem. What is needed is (1) a commitment of increase resources (chiefly in staffing) to the databanks; (2) a require-

ment by granting agencies that sequence data be submitted in machine-readable form to the databanks as a condition of funding; and (3) standardization of database design among the national databases.

In practice, each entry added to the genetic sequence databanks represents a loss of information; who would go through the 14,000-or-so literature citations and add the annotation to the sequences already present in the databanks? To prevent such loss of this information, the annotation and submission must become the researcher's responsibility.

DANIEL DAVISON

Department of Biochemical
and Biophysical Sciences,
University of Houston,
Houston, Texas 77004, USA

1. Pato, C.O. *Nature* 327, 467; 1987.

2. *GenBank Floppy Diskette User's Guide*, Release 48.0, February 1987.

The Universe as a fractal structure

Successful simulations do not validate the theories from which they spring, but it is remarkable that a numerical simulation of galaxy formation seems to reproduce the observed fractal pattern on the sky.

WHILE arguments persist about the mechanism by which galaxies were formed in the early Universe, cosmologists cannot complain that the issue is at risk of being hijacked by people with quite different interests. Even so, on the face of things it is a considerable surprise that one of the neatest models yet of the process of galaxy formation should have come from two people, both Hungarian, who have used the techniques of calculating with cellular automata to produce a persuasive simulation of the distribution of galaxies seen on the surface of the sky.

The essence of the problem, well-defined (notably by P.J.E. Peebles and his associates at Princeton) for at least a decade, is that while galaxies are on the average uniformly distributed through the volume of the Universe, the observed distribution of both optically visible and radio-galaxies on the sky is not uniform, but patchy. The observations show that galaxies tend to be clumped which has naturally provoked all kinds of speculation about the mechanism of their formation.

Does the clumpiness represent the distribution of matter at some primeval stage in the evolution of the Universe? Or have there been gravitational processes, Zeldovich's self-gravitating and fragmenting pancakes perhaps, which account for the present clumpiness of observable matter? Ostriker and Cowie have even suggested (*Astrophys. J.* **243**, 127; 1981) that the present distribution of galaxies is the relic of a dynamic process in which an outwards-propagating shock wave from an earlier generation of galaxies created galaxies at places of abnormally high density on the shock front.

What needs to be explained is the precise form of the correlation function, derived from the observations, representing the known clumpiness. The empirical rule is that the chance of finding a second galaxy within some volume unit at a distance s from a first is proportional to an inverse power of s . Qualitatively, this means simply that there is a greater chance that galaxies will be close together than far apart. Quantitatively, the value of the exponent in the inverse power-law comes out at 1.8. This is precisely the numerical value obtained by Tamas Vicsek and A.S. Szalay (*Phys. Rev. Lett.* **58**, 2828; 1987) with their model of the process of galaxy formation in which cellular automata simulate the outward spread of a shock-wave of the kind that is

suggested by Ostriker and Cowie.

Simple though this rule may seem, it has led directly to the recognition that the distribution of galaxies in the Universe may have a fractal (or fractional dimensional) structure. For what the rule implies is that, on the average, the number of galaxies in some volume centred on a particular galaxy is proportional to the dimensions of the region raised to the power 1.2 — the difference between the number of physical dimensions, 3, and the exponent of the correlation function, 1.8. This is exactly the kind of rule that applies, for example, to the formation of aggregates of particles diffusing towards a central nucleation site, and where the total number of particles in an aggregate scales not with an integral power of the dimensions, but with a non-integral or fractal power thereof.

By itself, of course, this argument does not say much about the way in which the observed large-scale structure of the Universe came to be, although Szalay and Vicsek cite their grounds for ruling out some processes for explaining the observed clumpy distribution of galaxies. For example, they say that self-gravitation is implausible because the galaxies now found in clumps should then, in general, be moving more quickly because of the conversion of potential to kinetic energy, which appears not to be the case.

A more direct argument for believing that the Universe has a fractal structure, however, is that presented by Szalay and another associate, D.N. Schramm, who argued two years ago (*Nature* **314**, 718; 1985) that the distribution of galaxies is mirrored by that of galaxy clusters; it seems that clusters also tend to cluster, and with an empirical distribution with the same numerical exponent. A large part of the interest of these arguments is that, if the structure of the Universe is indeed fractal, no special explanation will be needed for the observation that large tracts of the Universe appear devoid of galaxies; they will be just statistical flukes made probable by the clumping tendency.

So how to simulate this? Cellular automata are a way of simulating the hydrodynamics of a shock wave of the kind suggested by Ostriker and Cowie, for the time being intractable. The trick is to imagine that the Universe is represented by a cubic lattice. Each point of the lattice may be occupied or unoccupied by a galaxy and, in the simulation, a possible distribution of galaxies is obtained by supposing that

successive layers of points are added to small cubes centred on one corner of the whole structure.

As in all cellular automata simulations, there must be a rule for telling how the properties of the latest layer added to the structure are determined by the properties of its predecessor. The rule actually used supposes that the question whether each point in a newly added layer will (or will not) be occupied by a galaxy is mostly determined by the occupancy of the five nearest neighbours in the previous layer, but for good measure, there is a random variable to introduce an element of white noise to the system. To make the process a little more interesting, the determination whether a new site is occupied depends on whether a number characteristic of that site, and calculated by simple arithmetic from the corresponding number for the five nearest neighbours in the preceding layer, exceeds an arbitrarily chosen number.

The results are spectacularly successful. Most dramatically, the simulation does indeed yield a power-law relationship between the number of galaxies in a region around a central galaxy and the dimensions of that region, and with the expected exponent of 1.2. The simulation is almost exact over a 100-fold range of values of the length scale, breaking down only at the longest scales.

Even more surprising is the discovery that, when the results of the simulations (projected on a two-dimensional 'sky') are used to calculate the numbers of galaxies found in clusters of different sizes, the result is almost an inverse-square relationship (the exponent is 1.9, not 2.0) between the number of clusters of size s and the value of s , which is what seems to emerge from the real sky.

Vicsek and Szalay properly acknowledge that their success does not validate the model of galaxy formation that has guided their simulation. Even so, it will surely stimulate cosmologists to know that the presence of a galaxy at one place may be simply determined by the presence or absence of galaxies in the neighbourhood, even if the details of the Ostriker and Cowie mechanism are wide of the mark. Meanwhile, it will not have escaped the attention of people working in quite different fields that the authors have provided a valuable illustration of the usefulness of simulation by cellular automata.

John Maddox

Protein chemistry

Folding into the right shape

Robert Freedman

FORMATION of disulphide bonds, a characteristic feature of secreted and cell-surface proteins, has been known for some time to be an early post-translational event associated with protein folding in the lumen of the rough endoplasmic reticulum. Recent work by Roth and Pierce¹ and by Myllylä and Koivu² provides some finishing touches to our picture of how this occurs. In particular, these papers describe convincing evidence that the formation of disulphide bonds in protein folding at biosynthesis is catalysed by the enzyme protein disulphide-isomerase (PDI). At the same time, two reports by Evans *et al.* and by Lang *et al.* on pages 266 and 268 of this issue^{3,4}, respectively, reflect the growing interest in the significance of *cis-trans* isomerism of prolyl-peptide bonds in protein folding. One of these papers³ establishes the coexistence of such isomers within both the folded and denatured states of the enzyme staphyloccocal nuclease; the other⁴ confirms that the process of prolyl-peptide bond isomerization can be catalysed *in vitro* by enzymes.

Reduced unfolded proteins can be oxidized *in vitro* to form native-folded proteins with the 'correct' set of disulphide bonds, as in the classic experiments of Anfinsen (for example, ref. 5); but even when catalysed by a mixture of reduced and oxidized glutathione (or other thiol:disulphide couple) the process is slow, especially at pH values below 8.5. An enzyme capable of catalysing this process was discovered in the 1960s and shown to operate by catalysing thiol:protein-disulphide interchange with a broad protein substrate range. The enzyme (PDI) is found in the microsomal fraction of secretory tissues such as liver and pancreas, and hence it was proposed to be the catalyst of disulphide-bond formation during secretory protein biosynthesis. But this proposal was not universally accepted, and up to the late 1970s the need for an enzymic catalyst of protein-disulphide formation was challenged, and alternative functions were proposed for PDI, such as a role in insulin catabolism.

Nevertheless, it has gradually become clear that PDI is highly concentrated in cells active in the synthesis of disulphide-bonded proteins, that it increases in parallel with increases in synthesis of such proteins in systems as diverse as the developing chick embryo bone and the ripening wheat endosperm, and that it is located only within the lumen of the endoplasmic reticulum (and hence is a member of a small group of proteins of growing interest

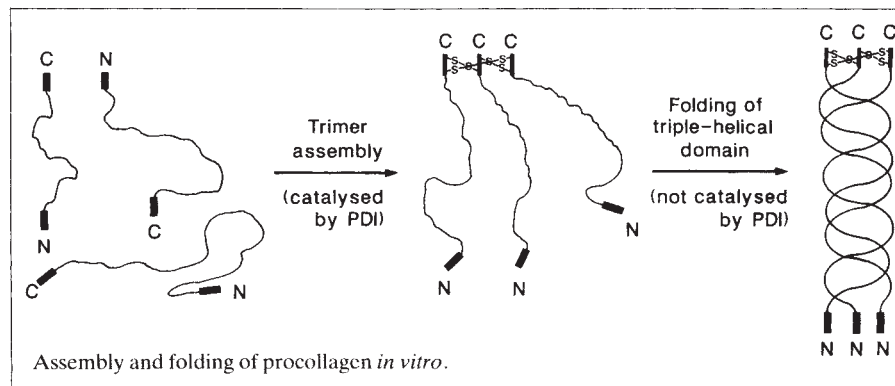
to cell biologists — but that is another story). All these indirect lines of evidence (for review see ref. 6) point to the enzyme being a catalyst for correct formation of disulphide bonds in protein biosynthesis, and in the past few years this idea has gained general acceptance.

The evidence, however, has never included a direct view of PDI at work in the cell, nor a clear demonstration with a complex multi-chain protein that PDI can catalyse correct disulphide formation and hence correct folding *in vitro* at the rate observed in the cell. The recently published work provides just that evidence.

Richard Roth and Marion Koshland demonstrated some years ago that the levels of PDI in various mouse lymphoid

the process of disulphide formation in the assembly of procollagen, a protein comprising three polypeptide chains linked by disulphide bonds at the carboxy termini and with globular regions at both ends that are cleaved in extracellular processing, leaving the central region which has the classical triple-helical collagen conformation (see figure). We have shown that PDI catalyses the formation of procollagen trimers from the free polypeptide chains *in vitro*, but we did not study the process in detail. Myllylä and Koivu have now studied² assembly of reduced procollagen chain into trimers and the formation of the collagen triple helix in the presence of PDI and of chemical catalysts of protein-disulphide interchange.

The results are striking. For type I procollagen, the optimal half-time for chemically catalysed trimer formation is 11 min, whereas in presence of PDI it is as short as 4 min; for triple-helix formation the half-times are 17 and 9 min, respectively. These results make two important points.



cell lines correlate closely with their activity in immunoglobulin synthesis. Now, Roth and Pierce¹ have tested the proposed role of PDI in immunoglobulin assembly, by studying what proteins are directly associated with PDI in intact lymphocytes. Their approach was to apply protein crosslinking agents to cells, solubilize the cells and then immunoprecipitate PDI with an anti-PDI monoclonal antibody; the immunoprecipitate could then be resolubilized, the crosslink cleaved and the associated proteins identified. Roth and Pierce's work shows clearly that newly synthesized immunoglobulin heavy and light chains can both be crosslinked to PDI in intact hybridoma cells which are actively synthesizing and secreting immunoglobulins. Several controls are necessary, for example to test for proteins co-precipitated with PDI when no chemical crosslinker is used, and for proteins precipitated by irrelevant antibodies, but the result is unambiguous. In cells synthesizing antibodies, significant amounts of antibody polypeptides are directly associated with PDI.

Raili Myllylä and Juha Koivu at the Collagen Research Unit at Oulu, Finland, took a different approach². They studied

First, the rates of trimer and triple-helix formation in the presence of PDI correspond to those found in biosynthetic studies in the cell; and second, the lag between trimer assembly and triple-helix formation is 5–6 min irrespective of the presence or absence of PDI. This lag is exactly that found in the cell, and the result is consistent with the view that the uncatalysed *cis-trans* isomerization of prolyl-peptide bonds in this proline-rich domain determines the rate of folding of the triple helix. The cellular significance of these results is confirmed by Koivu and Myllylä's studies² on type II procollagen. They found that it assembles very slowly in the absence of PDI *in vitro*, but in the presence of PDI it assembles with a half-time of 12 min and folds into the triple helix with a half-time of 17 min. The assembly and folding of type II procollagen is thus considerably slower than that of type I procollagen, and this is precisely the situation observed *in vivo*, where the slower folding of type II procollagen appears to allow more extensive hydroxylation of prolyl and lysyl residues.

The evidence thus converges from two directions. In the test-tube, PDI is capable of catalysing correct disulphide-bond

formation (and hence protein folding) in a complex multi-chain and multi-domain protein at a rate equal to that observed for the process in the cell. And in a cell that actively synthesizes and secretes a single major disulphide-bonded protein, PDI is in direct physical contact with that protein.

Detailed studies on the importance of prolyl-peptide bond isomerization to folding of proteins at biosynthesis are at a much more primitive stage. An extract capable of catalysing the *cis-trans* isomerization process in small synthetic proline-containing peptides was described in 1984, but has not yet been characterized extensively. Nevertheless, in their paper in this issue⁴, Lang *et al.* show that this preparation (peptidyl-prolyl *cis-trans* isomerase) can catalyse refolding of some denatured but correctly disulphide-bonded proteins (*cis-trans* isomerization of prolyl-peptide bonds is important in the folding of all these proteins). The catalysis appears to be enzymic, although high ratios of enzyme to protein substrate (1:5 and 1:20) are used to obtain significant catalysis. Considerable further work in characterizing the enzyme in terms of structure and action on a range of physiologically meaningful substrates, and in establishing its subcellular location, is required to determine whether this enzyme is involved in protein folding *in vivo*. But, without doubt, it will be useful in studies *in vitro* and the results now reported by Lang *et al.* emphasize that prolyl-peptide bond isomerization can be the rate-determining step in protein folding.

This is further highlighted by the detailed nuclear magnetic resonance (NMR) studies reported by Evans *et al.* in this issue⁵. These authors have been determining by NMR the structure in solution of the enzyme staphylococcal nuclease, and had noted⁶ that all the histidine residues in the native protein give rise to two sets of resonance signals, implying the co-existence of two distinct folded forms of the enzyme. They have now extended this observation in two ways. By observing the interconversions between the various states of the protein using the technique of magnetization transfer, they show that one histidine residue (His 121) exists in two alternative environments in the unfolded protein also. One of the only possible sources of these alternative environments for His 121 in the unfolded protein is *cis-trans* isomerism of the peptide bond before the nearby prolyl residue, Pro 117. In the crystalline state of the folded protein this bond has the *cis* conformation; the NMR results are explicable if the protein in solution exists as a mixture of the *cis* and *trans* isomers. In the unfolded protein, this isomerism would have local effects, so that only the nearest histidine residue has a significantly different environment in the two isomeric forms. But in the folded state, isomerism

could result in small changes in structure propagated throughout the protein, so that all the histidine residues essentially occupy slightly different environments in the two conformational isomers.

In addition, Evans *et al.*⁵ confirm the importance of the isomerism at Pro 117 by site-directed mutagenesis, in which they replace Pro 117 with a glycine residue that can have no such isomerism. The Pro 117 → Gly mutant protein is enzymically active and similar in conformation to the wild-type enzyme, on the basis of the broad features of its NMR spectrum. Strikingly, however, the NMR spectra of both the folded and unfolded mutant enzyme show only a single set of resonances for each histidine residue, implying that each is in a unique environment. Hence the presence of Pro 117 is clearly crucial for the alternative environments experienced by the histidine residues in the wild-type enzyme, and there seems

little doubt that it is the *cis-trans* peptide-bond isomerism of this residue which is significant. So, isomerization of prolyl-peptide bonds is not simply a process to be gone through in the initial folding of a protein. This isomerism can give rise to alternative, slowly interconverting states that can coexist in unfolded and folded proteins □

1. Roth, R. & Pierce, S. *Biochemistry* **26**, 4179-4182 (1987).
2. Myllylä, R. & Koivu, J. *J. biol. Chem.* **262**, 6159-6164 (1987).
3. Evans, P.A., Dobson, C.M., Kautz, R.A., Hatfull, G. & Fox, R.O. *Nature* **329**, 266-268 (1987).
4. Lang, K., Schmid, F.X. & Fischer, G. *Nature* **329**, 268-270 (1987).
5. Anfinsen, C.B. *Science* **181**, 223-230 (1973).
6. Freedman, R.B. *Trends biochem. Sci.* **9**, 438-441 (1984).
7. Forster, S.J. & Freedman, R.B. *Biochem. Rep.* **4**, 223-229 (1984).
8. Fox, R.O., Evans, P.A. & Dobson, C.M. *Nature* **320**, 192 (1986).

Robert Freedman is Reader in Protein Biochemistry at the University of Kent, Canterbury, Kent CT2 7NJ, UK.

Palaeomagnetism

Use and abuse of intensity data

Ronald T. Merrill

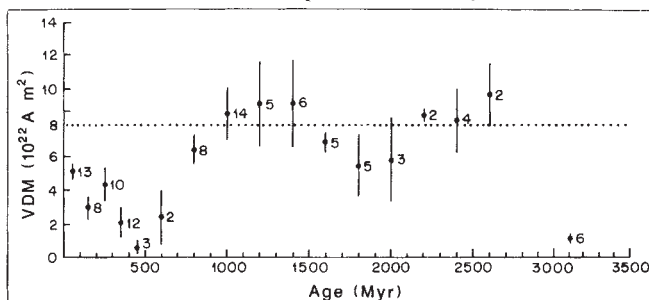
MAGNETIC directions of rocks can reveal interesting aspects of diverse problems, such as the history of the Earth's magnetic field, global tectonics, stratigraphy and palaeoclimate. The magnetic field is a vector quantity so that palaeointensity measurements could be as useful as palaeodirections¹. To date, however, only direction measurements have contributed to the most important advances in Earth sciences. C.J. Hale, on page 233 of this issue², reports new palaeointensity measurements that could indicate significant events in the Earth's history. His data show a sharp increase in the Earth's magnetic field between 2.7 and 2.1 gigayears (Gyr) ago. On the other hand, Walton challenged, in a recent issue of *Nature*³, the veracity of palaeointensity measurements in general.

Hale² argues that the sharp increase in

intensity that he detects arises from the nucleation of the Earth's inner core, such as some modellers suggest occurred between 2.5 and 1.5 Gyr ago⁴. He also notes that such events represent major changes in heat transfer in the Earth's interior that are ultimately revealed in changes on the surface. Hale sees evidence for such changes in other palaeomagnetic data that he interprets as showing important changes in the rate and style of continental drift at about 2.5 Gyr.

The figure shows some recent palaeointensity data⁵ together with Hale's six oldest values averaged and plotted as a single point at 3.1 Gyr. Although this figure does not do full justice to Hale's arguments, it does show the sharp increase in intensity before 2.5 Gyr and it also raises a problem. If one interprets the palaeointensity data as truly representing

the intensity of the Earth's magnetic field in the past, how does one interpret the marked dip around 0.5 Gyr? In the spirit of literal interpretation, perhaps this increase represents the time of inner-core nucleation, and the earlier increase identified by Hale represents some other event, for example, the growth of the core past some critical size. But are such



Variation of the mean virtual dipole moment (VDM) with time (from ref. 5). 100 Myr averages are used from 0 to 700 Myr, and 200 Myr averages from 700 to 2,700 Myr. Hale's six oldest palaeo-intensities, which span nearly 800 Myr, are plotted at 3,100 Myr. The number of measurements is shown next to each mean, and standard errors are indicated. Dotted line: strength of the present-day dipole moment.

literal interpretations of the data justified? A more conventional and conservative interpretation of the palaeointensity data is that the magnetic-field intensity has always been similar to that of the present field. An even more conservative interpretation, held by some in the field, is that all the data used in the above arguments are suspect.

To understand why such diverse views exist, some of the difficulties involved in obtaining reliable palaeointensity estimates must be understood. The present-day magnetic field is regatively complex. Only 90 per cent of the field at the Earth's surface can be described as resulting from a geocentric dipole tilted at 11° to the rotation axis. Also, the field at the Earth's surface can change significantly in both direction and intensity in just a few tens of years.

To obtain an average palaeointensity estimate, palaeomagnetists use the substantiated approximation that the time-averaged magnetic field at the Earth's surface can be represented as originating from a geocentric dipole. This allows the magnetic-field intensity to be characterized by a single number (such as the dipole moment). Even in this case, there are good reasons for believing that the average dipole intensity can change significantly over a few hundreds to a few thousands of years⁵. Because even the best radiometric dating is rarely accurate to more than 2 or 3 per cent, and because the data are few, there may be significant biases in the palaeointensity trends shown in the figure.

According to Walton³ in his recent papers an even more serious problem involves potential errors in palaeointensity estimates obtained with rocks that undergo chemical alteration. Absolute palaeointensity estimates are made from rocks and archaeological artefacts that are believed to have acquired a thermal remanent magnetization on cooling from high temperatures in the Earth's field. This remanent magnetization, M , is usually linearly proportional to the weak inducing field of the Earth, H ; that is, $M = CH$, where C is a constant of proportionality. To obtain an estimate of H , C must be determined first. This can be done by reheating the sample and cooling it in a known laboratory field, H_L , so that a new remanent magnetization $M_L = CH_L$ is induced. Obviously, this gives C , so that the estimate of H can be obtained.

As is well recognized, however, M is not usually a purely thermal remanent magnetization, and C does not usually remain invariant from the time the sample formed until it has been reheated. Chemical alteration, viscous effects, deformation and other processes, for example, can lead to changes in M and C . Consequently, the purpose of all palaeointensity techniques is to obtain suitable samples from which

several independent estimates of H can be obtained from the same sample. It is usually argued that if these estimates all agree, the value is reliable. The detailed methods for determining these independent estimates are almost as numerous as the investigators working on this problem, but essentially all techniques use the observation that the remanent magnetization of a sample typically consists of many remanent magnetizations that exhibit different coercivities and thermal properties; independent estimates of H are determined from these several different magnetizations.

Walton³ argues that the checks in one of the most commonly used techniques, a modified version of Thellier's technique, fail to detect the effects of thermally activated chemical alteration. The argument that chemical effects can lead to erroneous palaeointensity estimates has been made before by Walton and by others, but the adverse effects were thought to be detectable, except in very rare cases. (More conventional interpretations are given in refs 6 and 7.)

Walton combines theoretical arguments and experimental data to argue that erroneous palaeointensity estimates may be common, and provides a method to correct for chemical alteration that can occur during laboratory reheating. As he acknowledges, however, this method will

not detect adverse effects associated with any previous alteration. Because Walton has not identified the mineralogical changes that occurred in his samples, it is possible that alteration could also have taken place during the heating and cooling that occurred before the sample was brought to the laboratory. Consequently, his method for obtaining reliable palaeointensity estimates also appears to be flawed.

Given the reception at scientific meetings to Walton's views, one can anticipate that his newest paper will also be controversial and unpopular with some palaeomagnetists. Nevertheless, given the great implications that reliable palaeointensity estimates can have in the Earth sciences, as shown by Hale's paper in this issue, it is important to consider carefully Walton's challenge that most, and perhaps all, palaeointensity estimates are suspect. □

1. Thellier, E. & Thellier, O. *Ann. Geophys.* **15**, 285–376 (1959).
2. Hale, C. *Nature* **329**, 233–237 (1987).
3. Walton, D. *Nature* **328**, 789–791 (1987).
4. Stevenson, D. *Rep. Prog. Phys.* **46**, 555–620 (1983).
5. Merrill, R. & McElhinny, M. *The Earth's Magnetic Field* (Academic, London, 1983).
6. Coe, R. *et al.* *J. Geomagn. Geoelect.* **25**, 415–435 (1973).
7. Aitken, M. *et al.* *Archaeometry* **23**, 53–64 (1981).

Ronald T. Merrill is at the Department of Geophysics, AK-50, University of Washington, Seattle, Washington 98195, USA.

Evolutionary biology

Environmental determination of sex in the reptiles

Graham Head, Robert M. May and Linwood Pendleton

GENETIC sex determination predominates among vertebrate species. Nevertheless, in many reptiles and in some other taxa, sex is determined by environmental factors, such as temperature, that act after conception. In the past ten years, a good deal has been learned about the adaptive significance of such environmental sex determination, about the cues involved, and about when they act^{1–3}.

There remains some uncertainty, however, as to why the precise pattern of environmental sex determination (ESD) varies so widely within certain taxa, particularly reptiles^{1,4}. In most turtle species, for example, only females are produced at high temperatures and only males emerge at low temperatures. The converse is true in many crocodilians (crocodiles and alligators) and in some lizards. In other species of lizards, and probably in all snakes, sex is genotypically determined¹. The figure summarizes this variety of observed patterns, and poses the question of why this should be so. To find an answer, one must understand why ESD

should arise at all.

ESD will be favoured under the following two conditions. The first is that individuals be lastingly affected by the environment they encounter early in life, and that this environment be spatially patchy, with some patches increasing (or decreasing) male fitness more than female fitness, and with the converse being true in other patches. The second condition is that neither parents nor offspring can control which type of patch offspring enter. Under these circumstances, genetic sex determination (GSD) would lead to individuals often having the sex of lower fitness for their particular environment. ESD, on the contrary, allows sex ratios to be adjusted in response to specific environmental conditions⁵.

For reptiles, the environmental factor determining sex is temperature (although water potential may also be involved¹). Higher temperatures mean faster growth rates, shorter incubation periods, larger juvenile sizes, and hence larger adult sizes⁶. Thus, if the benefits derived from

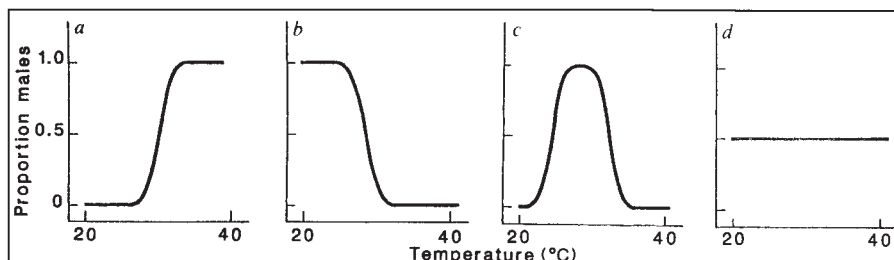
large adult size differ between the sexes, that sex which benefits more from being relatively large should be the one that develops at higher temperatures. This suggests that variation in the pattern of ESD is a consequence of males gaining a greater advantage from being large in some reptile species, while females benefit more in others. If so, there should be a direct correspondence between the patterns of ESD and an index, such as the patterns of sexual dimorphism in size, that can be used to indicate which sex gains the greater advantage from large size.

The trends in sexual dimorphism in size among the different reptilian orders show that in most species of turtles, females are larger than males, owing to the relationship between female size and fecundity^{7,8}. As predicted, when ESD occurs in turtles, females are almost always produced at warmer temperatures, where growth rates are higher (*b* in the figure). On the other hand, male lizards and crocodilians are usually larger than females, reflecting the importance of large size in male–male competition for mates in these orders⁸. In these groups, males are produced at warmer temperatures (*a* in the figure). Hence, the general patterns of ESD appear to be explicable in terms of trends in sexual dimorphism.

What, then, is the significance of *c* in the figure? In this form of ESD, males are produced only at a limited set of intermediate temperatures, which may be optimal temperatures for embryo growth. The phylogeny of taxa in which this pattern is present suggests it may be derived from the pattern of *a*. Possibly the pattern of *c* occurs in species whose eggs must face wide and unpredictable ranges of temperatures, either because of a large geographical range (for example, the painted turtle, *Chrysemys picta*), or because of a large clutch size that results in a temperature gradient within a nest (for example the snapping turtle, *Chelydra serpentina*¹⁰). Under these circumstances, this type of ESD would ensure that a reasonable proportion of females are produced in all environments.

There remain the species of reptiles in which sex is genotypically determined. All snakes, many lizards and some turtles exhibit GSD. These are, for the most part, the reptilian groups that have evolved relatively recently. The presence or absence of ESD thus may be related to phylogeny, with GSD possibly being a derived state that dominates in modern orders and occurs in some genera within older orders¹. Why this should be so is not yet understood.

Those reptile species that are exceptions to the general patterns of ESD must also be explained. Snapping turtles provide one example. Unlike other turtles, they show the form of ESD exhibited by crocodiles (*c* rather than *b*). However,



Response of sex ratio to incubation temperature among reptile species. These schematic curves represent only the approximate form of the response, and do not represent the actual data for any one species. Four patterns are recognized: *a*, females develop at low temperature, males at high temperature, as in two lizards and in alligators; *b*, the reverse of *a*, males develop at low temperatures and females at high ones, as in most turtles that have been studied; *c*, females at low and high temperatures, males at intermediate ones (this pattern may be widespread, occurring in snapping turtles and perhaps other turtles, and in an Australian crocodile); *d*, the ratio of some species is not significantly influenced by incubation temperature. (After ref. 1.)

also unlike most other turtles, male snapping turtles are larger than females and are highly aggressive⁷. In this sense, they resemble crocodilians rather than 'typical' turtles. Also, they have large clutches of eggs, as do crocodilians⁸. Other examples can be understood in similar terms; species showing the reverse of expected ESD patterns also usually have their patterns of size dimorphism reversed.

Several problems remain. ESD data are hard to collect, and some sets of data are not as complete as they might ideally be. In particular, distinguishing between *a* and *c* is difficult, for various technical reasons; it seems that most crocodilians, and many turtles, may have been too simply assigned to *a*, when in fact they exhibit the pattern of *c*. Perhaps the greatest problem, however, lies in the conjectured link between higher temperatures and increased fitness of offspring. Some evidence suggests that lower temperatures may, on occasion, actually lead to higher fitness. For example, in the American alligator, *Alligator mississippiensis*, eggs laid at lower temperatures have more yolk, which enables more rapid growth; as a result, juvenile females are larger than juvenile males, which may increase the survival of females¹¹. Why more yolk should not be deposited at higher temperatures is unclear. Other factors, such as females of a given size selecting nest sites of a given temperature, may be involved. Whatever the reason for this result, more data are needed on the incubation periods and growth rates of juvenile males and females of the different reptilian orders.

Patterns of ESD have broader implications for reptilian biology. Reptilian sex ratios will in general be affected by the form of ESD. Such sex-ratio effects will, however, often interweave with other factors, such as local mate competition. Thus if the proportions of males in clutches of each of six freshwater turtle species (*Clemmys marmorata*, *Pseudemys scripta*, *Chelydra serpentina*, *Kinosternon subrubrum*, *K. flavescens* and *Sternotherus odoratus*) are plotted against the

size of the home range — which presumably reflects the potential for local mate competition — a positive correlation is found⁶; relatively fewer males are produced in those species where local mate competition would otherwise be greatest. The presence of local mate competition therefore confounds the analysis, making it difficult to determine whether the pattern of ESD has any consistent effect on sex ratios.

A more general question, which has received essentially no systematic exploration, concerns the effects of sustained and significant changes in global or local climate upon creatures with the kinds of ESD shown in *a* and *b*. If the range of temperatures demarking the development of males from females is sufficiently inflexible, a large-enough change in average temperature could effectively extinguish one or other sex, and thence the species. Indeed, there has been speculation, necessarily unsupported, that climate changes around the Cretaceous–Tertiary boundary extinguished the dinosaurs by this mechanism. Less dramatically, such effects of systematic climate change could produce small effective population sizes of either males or females, producing genetic 'bottlenecks', along with all the attendant evolutionary consequences^{12,13}. □

1. Bull, J.J. *Evolution of Sex Determining Mechanisms* (Benjamin/Cummings, Menlo Park, 1983).
2. Bull, J.J. & Vogt, R.C. *J. exp. Zool.* **218**, 435–440 (1981).
3. Conover, D.O. & Heins, S.W. *Nature* **326**, 496–498 (1987).
4. Bulmer, M. *Nature* **326**, 440–441 (1987).
5. Charnov, E.L. & Bull, J.J. *Nature* **266**, 828–830 (1977).
6. Harless, M. & Morlock, H. (eds) *Turtles: Perspectives and Research* (Wiley, New York, 1978).
7. Berry, J.F. & Shine, R. *Oecologia* **44**, 185–191 (1980).
8. Fitch, H.S. *Univ. Kansas Mus. Nat. Hist. Misc. Pub.* **70**, 1–72 (1981).
9. Bull, J.J., Vogt, R.C. & McCoy, C.J. *Evolution* **36**, 326–332 (1982).
10. Wilhoft, D.C., Hotelling, E. & Franks, P. *J. Herpetol.* **17**, 38–42 (1983).
11. Ferguson, M.W.J. & Joanen, T. *Nature* **296**, 850–853 (1982).
12. Frankel, O.H. & Soule, M.E. *Conservation and Evolution* (Cambridge University Press, 1981).
13. Bryant, E.H., McCommas, S.A. & Combs, L.M. *Genetics* **114**, 1191–1211 (1986).

Graham Head, Robert M. May and Linwood Pendleton are in the Biology Department at Princeton University, Princeton, New Jersey 08544, USA.

Archaeology

New dates for the Acheulean age

J.A.J. Gowlett

THE timescale of the Old Stone Age was completely changed more than 20 years ago by potassium-argon dates¹ that showed that the early archaeological and hominid sites at Olduvai Gorge were more than twice as old as had been generally recognized. These dates, of about 1.8 million years (Myr) before present, left the subsequent million years of Stone-Age prehistory unexplained. The new dating of the Acheulean sites of Olorgesailie in Kenya reported by Bye *et al.* on page 237 of this issue², which now jump from about 0.4 to 0.7–0.9 Myr, can be seen as one of the last aftershocks in the process of readjustment.

The Olorgesailie complex, discovered by the Leakeys in the 1940s among Pleistocene lake beds in the Rift Valley of Kenya³ has remained one of the principal exemplars of the Acheulean tradition, largely because of the extensive excavations and detailed studies by the late Glynn Isaac⁴. The Acheulean is the longest Stone-Age tradition, found all over Africa and in much of Europe and Asia. It has many facies, but above all is characterized by large, skilfully shaped stone tools called bifaces or hand axes. Its overall chronological range is reasonably well known^{5,6} — from about 1.4 to 0.15 Myr — but within that span there have been few fixed points, and so it has been difficult to examine rates of cultural change.

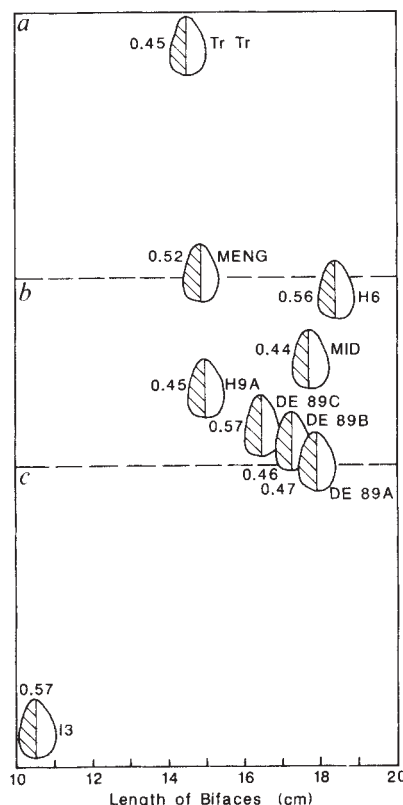
Although Isaac was quick to explore the implications of the longer timescale⁷, Olorgesailie, his first major site, fell squarely in what he cheerfully termed “the muddle in the middle”. Irritating as this lack of a firm date must have been, it undoubtedly helped to shape his contribution to archaeological theory, in particular the exploration of problems of cultural (artefact) variation in circumstances where time relationships were poorly defined. Previous work on other important sites in eastern Africa, particularly Kalambo Falls⁸ and Isimila¹⁰, set the scene for this approach. Traditionally, the Old Stone Age had been seen as a period of slow but steady progress, so that an artefact might be dated approximately by its level of sophistication.

The variation seen among contemporary archaeological occurrences within a site complex, as at Isimila, contradicted this view. Isaac concluded that some aspects of assemblage variation at Olorgesailie, in frequency of tool types, for example, were likely to be functional, reflecting different activities. Others, such as variations in hand-axe shape and size, he attributed to a random drift of style. There was sufficient

flexibility in a craft tradition to allow cumulative changes, perhaps over several generations, but sufficiently rigorous general limits to keep hand axes to the same pattern over far longer periods.

The new potassium-argon dating described in this issue² does not contradict this interpretation. It has the effect of placing most of the East African Acheulean within the Lower Pleistocene, older than 0.7 Myr. Olorgesailie now joins Peninj⁹, Olduvai Bed IV⁶, Kilombe and Kariandusi¹¹ in the earlier half of the Acheulean realm. Although stylistic grounds are not dating evidence, much of the Olorgesailie material sufficiently resembles Olduvai Bed IV and Kilombe, that its approximate contemporaneity now demonstrated by Bye *et al.* reinforces the idea that there was a consistency of Acheulean style within certain periods.

Most of the Olorgesailie localities are almost contemporaneous, but Isaac suspected that the artefacts from member 1 (the lowest stratigraphic layer) might be considerably older. This may be supported



Variation among bifaces from the sites at Olorgesailie in Kenya (data from ref. 4). Means of the thickness-breadth ratio are indicated to the left of the symbols. a, Upper stratigraphic set, about 10 metres sediment. b, Middle set, about 1 metre cut-and-fill sediment. c, Lower set, about 10 metres sediment.

by the new evidence that members 1–4 have a different tuff composition from those above. In relation to Site 13 in member 1, Isaac noted that the larger bifacial tools from members 1 and 2 contrast markedly with those from all higher strata. The modal forms of the lower stratigraphic set are smaller and relatively thicker and the trimming scars are characteristically large and deep, and core-like bifaces are common. This could indicate a significant age difference, but not necessarily a straightforward evolutionary progression, as yet earlier bifaces from Olduvai and Peninj exhibit different features. Better knowledge of time relationships in the Acheulean is at last helping to map out the field of variation.

The other principal aspect of the Olorgesailie complex is the preservation of bones on some sites. Meat was an important item of food at some sites, such as DE/89, where many bones and teeth of the giant baboon *Theropithecus* were found with the artefacts. In the 1960s it was common to interpret this association as evidence of skilled hunting¹², but some specialists now attribute it to scavenging and/or the effects of taphonomy — nature's tricks of deposition. The issue is not yet settled.

From his view of the Olorgesailie evidence, Isaac concluded that during this period “the last crucial evolutionary changes that culminated in mankind as we know it, took place”. But he confessed that the period lacked the glamour of searching for human origins, adding ruefully that many aspects of the Middle Pleistocene record “strike even enthusiasts as monotonous”. Olorgesailie has a central place in the evidence of the period, and is now exciting again, with the new dates taking it out of the Middle Pleistocene and into the late Lower Pleistocene. We are left with a new problem: what important African sites can be placed with certainty in the period 700,000–300,000 years ago? □

1. Evernden, J.F. & Curtis, G.H. *Curr. Anthropol.* **6**, 343–385 (1965).
2. Bye, B.A., Brown, F.H., Cerling, T.E. & McDougall, I. *Nature* **329**, 237–239 (1987).
3. Leakey, L.S.B. *Proc. 1st Panafr. Congr. Prehist.* (eds Leakey, L.S.B. & Cole, S.) 209 (Blackwell, Oxford, 1947).
4. Isaac, G.L. *Olorgesailie: Archaeological Studies of a Middle Pleistocene Lake Basin in Kenya* (University of Chicago Press, 1977).
5. Isaac, G.L. & Curtis, G.H. *Nature* **249**, 624–627 (1974).
6. Leakey, M.D. in *After the Australopithecines* (eds Butzer, K.W. & Isaac, G.L.) 477–494 (Mouton, The Hague, 1975).
7. Isaac, G.L. *Wild Archaeology* **1**, 1–28 (1969).
8. Isaac, G.L. in *After the Australopithecines* (eds Butzer, K.W. & Isaac, G.L.) 875–877 (Mouton, The Hague, 1975).
9. Clark, J.D. *Kalambo Falls Prehistoric Site*, Vol. 1 (Cambridge University Press, 1969).
10. Howell, F.C., Cole, G.H. & Kleindienst, M.R. in *Actes du IVème Congr. Panafr. Prehist.* (eds Mortelmans, G. & Nenquin, J.) 43–80 (Musée Royal de l'Afrique Central, Tervuren, 1967).
11. Gowlett, J.A.J. in *Proc. 8th Panafr. Congr. Prehist. Quat. Stud.* (eds Leakey, R.E. & Ogot, B.A.) 213–217 (TILMIAP, Nairobi, 1980).
12. Isaac, G.L. in *Man the Hunter* (eds Lee, R.B. & Devore, I.) 253–261 (Aldine, Chicago, 1968).

J.A.J. Gowlett is in the Institute of Prehistoric Sciences and Archaeology, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK.

Crystallography

Entering a new phase

B. W. Batterman

CRYSTALLOGRAPHERS recording the diffraction patterns of X-rays scattered by crystals are losing vital information — the phase of the X-rays. On page 232 of this issue, Shen and Colella offer a practical solution to the phase problem. If it can be solved, then the intensity of the Bragg reflections, which appear as spots on diffraction patterns, can be used to determine unambiguously the position and type of all the atoms that make up the fundamental unit from which the crystal is built.

The intensity of a wave producing the Bragg reflection is proportional to the square of the amplitude of its sinusoidal oscillation. When an X-ray wave strikes the film, the blackening produced is proportional to the intensity of the wave, and from this the amplitude can easily be determined. The phase of the sine wave (where it goes through zero in time with respect to the other Bragg spots) is lost, and to reconstruct the three-dimensional crystal from its two-dimensional diffraction pattern the phase and the amplitude of each Bragg reflection are needed.

Analogous to the phase problem is the difference between the holographic image of an object and its standard camera representation on a sheet of film. The film image clearly lacks information available from the hologram. The latter is a full three-dimensional representation of the object, so that one can see all sides of it; the former gives the same limited two-dimensional information irrespective of how it is viewed. The three-dimensional character of the hologram derives from the fact that the relative phase of the light from each point on the object is recorded.

The problem is how to record intensity and phase information on film that is sensitive only to intensity. To do this requires changing the phase information into intensity information, achieved by comparing the scattered waves with an unscattered reference wave. The interference pattern obtained by mixing the two travelling waves is a series of peaks and troughs that are stationary in time producing a moiré-like pattern of fringes whose intensities depend on the amplitude of the scattered wave and its phase relative to the reference. Moving the object changes the phase, so that the pattern of fringes will change. The intensity and phase of each object point are simultaneously recorded by the film.

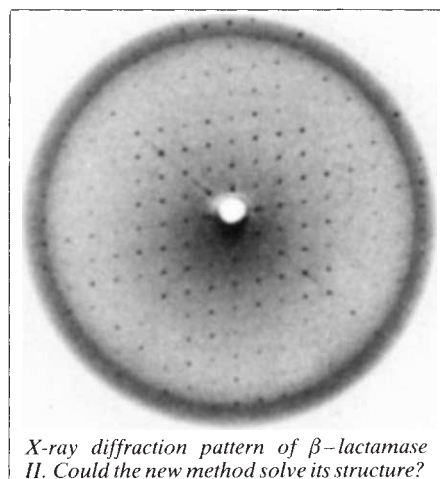
A more conventional approach to phasing is a technique called isomorphous replacement. In this case, the reference phase is obtained from the scattering of a heavy atom inserted in the structure,

preferably with a minimal perturbation of the surrounding atomic positions. Isomorphous replacement is an invasive technique with limited applicability.

Statistical phasing of the direct method variety was pioneered by Karle and Hauptmann who were awarded the 1986 Nobel prize for their achievements. Their method takes account of certain relationships between phases of reflection derived from the global condition that the electron density must be everywhere positive. Although a powerful technique for small molecule structures it has not been successfully applied to large proteins.

For X-ray diffraction, Shen and Colella transform the phase of a particular Bragg reflection to an intensity variation by simultaneously exciting a second Bragg reflection that acts as a reference wave. Multiple Bragg excitations have been used by others (Colella, R. *Acta crystallogr.* **A30**, 413; 1974 and Post, B. *Phys. Rev. Lett.* **39**, 760; 1977) to obtain the phase of a given reflection. In general only crystals of relatively simple structure and extremely high perfection have been used to demonstrate the ability to measure phase. Few crystals are highly perfect and the most interesting have complex structures.

Shen and Colella show that even with crystals of poor perfection, multiple-beam excitation can be used to obtain the relative phase of a given Bragg reflection with respect to a reference reflection. Their method, demonstrated with a crystal of benzil, involves the rather straightforward



X-ray diffraction pattern of β -lactamase II. Could the new method solve its structure?

Brian Sutton

recording of the intensity of the reflection as the reference beam is brought into Bragg diffracting conditions. The observed intensity displays an asymmetrical peaking as the reference beam is continuously swept through its Bragg condition while maintaining Bragg reflection of the primary beam. The phase of the reflection is found by measuring the relative intensity difference between the left and right wings of the asymmetrical peak.

Shen and Colella have solved the phase problem using a relatively simple technique, the limits of which will no doubt be pursued by the crystallographers. Key questions to be answered are: can a sufficient number of reflections be phased to solve any particular structure? How imperfect can the crystal be? What is the upper limit to the size of the unit cell for which the technique is effective? The answers will prove the efficacy of Shen and Colella's technique. □

B. W. Batterman is at the Cornell High Energy Synchrotron Source, Cornell University, Ithaca, New York 14853, USA.

Mathematics

Geometry goes non-commutative

John D.S. Jones

A VIEW sometimes expressed is that the basic objects of study in mathematics are numbers, geometry and physical phenomena, and that the basic methods with which one studies these are algebra and analysis. Most mathematicians would greet this assertion with something like "Yes, but what about. . . ." Nonetheless there is an element of truth in the idea. Recently, Alain Connes has performed the mathematical equivalent of the three-card trick: he has found a rather unexpected way of applying some sophisticated parts of modern analysis, which have their origins in mathematical physics, to geometrical problems (*Pub. Math. IHES* **62**, 41–144; 1985). One exciting aspect of his approach is the way he draws on

experience and intuition from different areas of mathematics and makes links between them. Ideas from topology and geometry are mixed with ideas from analysis and mathematical physics to the benefit of all. It is intriguing that one of the things to come out of this mixture is the introduction of a new algebraic technique, cyclic cohomology.

To anyone with a little mathematical knowledge, the title of Connes' articles (*Non-Commutative Differential Geometry I and II*) may sound ridiculous. After all, the commutative law is the rule of multiplication, $ab = ba$. There are algebraic systems that do not obey this law, for example the multiplication of matrices, but at first sight such systems do not seem

to have much relevance to geometry. A key idea behind Connes' work is that geometrical problems can be modelled by an algebraic structure in which the commutative law does not hold. This structure contains useful information about the original problem, and this information needs to be extracted and interpreted in terms of the original geometry.

To understand some of these ideas it is necessary to go back several steps. It is a well-established and important mathematical idea that geometrical objects can be modelled by commutative algebraic structures. An algebraic system with an addition and a multiplication satisfying the normal rules, except the commutative law for multiplication, is called an algebra; if the multiplication satisfies the commutative law the algebra is called commutative; if not, non-commutative. Now suppose we are given some space which in general will have a high dimension, and be geometrically or topologically complicated. The algebraic structure associated to the space is the collection of all the functions on it. To avoid obtaining a trivial theory, we concentrate on continuous functions, and we can get more refined information by considering more refined classes of functions. Functions can be added and multiplied, and the multiplication of functions is necessarily commutative, so that the functions on the space form a commutative algebra. The idea that the algebra of functions on a space captures much of its geometry is at the heart of the development of general algebraic geometry by Alexander Grothendieck and his school in the 1960s and early 1970s. It gives a powerful and flexible way of approaching problems in geometry and makes a significant connection between geometry and number theory (see, for example, Ian Stewart's

News and Views article on algebraic number theory in *Nature* **310**, 729; 1984).

There are geometries, however, for which the natural algebra of functions contains no information. It often happens, for example, that the space has symmetry, so that a group of transformations acts on it. Obviously, this symmetry should be taken into account when studying the geometrical properties of the space. The natural way to do this is to look only at those functions which take the same value at two points that can be transformed to each other by one of the symmetries. Such functions are called invariant or symmetrical. There are simple examples in which the only invariant functions are constant. Three important points should be made: examples of this kind exist; they are common; and, most important of all, they are in no way weird and unnatural. The one I describe below involves nothing more than rotations of the circle, one of the most natural of mathematical objects.

In this example, the circle is the space and the group of transformations consists of all rotations through integer multiples of a fixed angle. If this angle is irrational then the set of points obtained by applying one of these rotations to a given point is dense in the circle, approaching arbitrarily close to every point on the circle. The invariant function is constant on this dense set, and the requirement of continuity then implies that it is constant on the whole circle. Note that this difficulty disappears if the chosen angle is rational, because then the set of points obtained from a given point is finite.

These examples show that in general we cannot expect to get enough information from the algebra of invariant functions. The way to get algebraic models in this case is to suspend the commutative law. It is possible to construct an algebra which is non-commutative but provides an acceptable substitute for the algebra of invariant functions. The awkward geometrical fact that there are no invariant functions is replaced by the (rather less) awkward algebraic fact that the commutative law is not obeyed.

What kind of algebras do these constructions produce and what is known about them? To answer this question we need to delve into the theory of operator algebras. These algebras were introduced in the 1930s in a series of papers by Francis Murray and John von Neumann, who were motivated by questions in quantum theory. The subject has come a long way since then but until relatively recently it was not thought to have any serious geometrical applications. The primary concern was to provide rigorous analytical techniques for studying subjects like quantum theory, group representations and statistical mechanics. One nice feature of the construction is that it follows one of the working principles of quantum

theory — replace functions by operators.

The next question to consider is how to process these algebras to get useful geometrical information. Here we find ourselves in the technical heart of the subject, but I shall try to illustrate the ideas by means of integration. Geometrically speaking, integration is concerned with calculating quantities such as area and volume, but Connes distils it to the mathematics of expressions $I(a_0, a_1, \dots, a_n)$ with two characteristic properties:

$$I(a_0, a_1, \dots, a_n) = (-1)^n I(a_n, a_0, \dots, a_{n-1})$$

$$I(a_0 a_1 a_2, \dots, a_{n+1}) - I(a_0, a_1 a_2, a_3, \dots, a_{n+1}) \\ + \dots \pm I(a_0, \dots, a_n a_{n+1}) \\ \pm I(a_{n+1} a_0, a_1, \dots, a_n) = 0$$

The signs in the second equation alternate and the a s are taken from the relevant algebra. Roughly speaking the a s play the role of functions and the expressions $I(a_0, \dots, a_n)$ are abstract models for certain multiple integrals. The two basic formulae are derived from the change of variables formula and integration by parts. These equations express basic properties of integration in a very abstract form; it is quite remarkable that they are sufficient to retain the technical power of integration yet flexible enough to allow Connes to develop his ideas in this non-commutative framework.

The expressions I are called cyclic cocycles, and from these Connes builds the theory of cyclic cohomology. This theory is a substitute in this non-commutative geometry for the theory of integration but in addition it provides a useful source of geometrical and topological invariants for the underlying geometrical object. One feature of the end product is that although the motivation was to extract geometrical information by using operator algebras, the whole thing can be turned on its head and the new theory can now be used to give information about algebras.

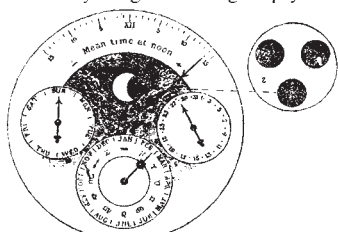
So this is the scheme of Connes' non-commutative geometry: model geometrical situations by non-commutative operator algebras and ask how to extract geometrical information from these algebras. Clearly, this is a very abstract general idea, but it works. It gives concrete information in the motivating examples and it provides a new algebraic technique. It not only answers questions, but the questions asked force the development of new techniques and quickly lead to a surprising combination of different parts of mathematics; one can hardly ask for more! It is everything that mathematics should be. $\square$

John D.S. Jones is at the Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK.

100 years ago

PROGRESS IN CLOCKS AND WATCHES

The form of mechanism for the purpose of maintaining and showing a calendar has undergone considerable development. Calendars are now made to be perpetual, correcting themselves for everything, including leap-year.



Until three years ago there was no public institution in Great Britain where a serviceable authentic trial of the performance of a watch under varying conditions could be obtained. At that date, however, under the auspices of the Royal Society, a department of the Kew Observatory was established for the purpose.

From *Nature* **36**, 486; 22 September 1887.

Dwarf galaxies

Icebergs and crouching giants

Michael Disney and Steven Phillipps

THE recent serendipitous discovery by Bothun *et al.*¹ that an apparently insignificant dwarf in the Virgo Cluster is in fact a huge, low-surface-brightness galaxy in the background has dramatically demonstrated that new populations of galaxies remain to be detected. This particular 'iceberg' (with a central disk surface brightness in the blue of 26.5 magnitudes arc s⁻², almost all of its light lies below the detection threshold of normal observations) was detected by Malin's method of photographic amplification and is known as Malin 1 (ref. 1). It is indisputably a 'crouching giant', for despite its apparent insignificance, its probable disk scale length of 55 kiloparsec and total neutral-hydrogen (H I) mass of 10¹¹ solar masses, make it certainly the largest and the most H I-rich disk known and also one of the brightest of all spiral galaxies¹.

Although Malin 1 could be a unique case there is in fact a significant body of circumstantial evidence to suggest otherwise. Our present knowledge of the galaxy population is so biased by a single insidious selection effect that it is entirely possible that Malin 1 is just the first example of a class of such low-surface-brightness giant galaxies that forms a significant constituent of the Universe.

When we compared³ the photometric properties of several hundred catalogued disk galaxies, we found that they all have essentially the same surface brightness, despite differences in luminosity, colour and morphological type (see Fig. 1). Also the common value of central surface brightness chosen by catalogued disk galaxies is just such as to give them the largest apparent size when seen against the sky background on a photographic plate. The apparent size of a galaxy of fixed intrinsic luminosity will depend on its surface brightness Σ ; for high Σ , the galaxy will

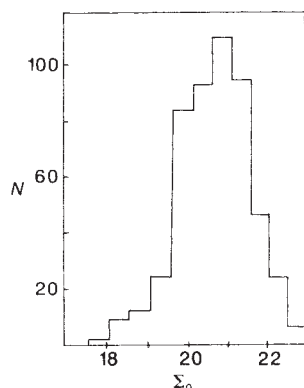


Fig. 1 Distribution of characteristic surface brightnesses for all well-studied disk galaxies in the RC2 catalogue (from ref. 3).

be intrinsically compact, but if Σ is low, most of the galaxy will be undetectable. Observed disks have Σ close to an optimum value between these extremes⁴, giving a maximum apparent size. Also the observed surface brightness of disk galaxies is close to the maximum allowed by internal absorption by dust⁵.

Elliptical galaxies with a wide range of intrinsic luminosities also have closely similar surface brightness⁶ (although different from that of spirals). Again, the common value of Σ for catalogued giant ellipticals is just such as to give them the largest apparent size. It is as if ellipticals, like spirals, make themselves as conspicuous as possible to contemporary Earth-bound observers.

Such a series of coincidences³ presents us with an intriguing challenge; it may be the Rosetta stone needed to decipher the physics of galaxies and their formation, or it may be nothing more than the inevitable outcome of observational selection, in which case there are many more galaxies about than we are presently aware of.

The night sky is, astronomically speaking, exceedingly bright. Indeed with the simple detectors so far available for wide field optical surveys it may be something of a surprise that we can see anything much of the extragalactic Universe at all. Our view of a low-surface-brightness nebulous object such as a galaxy clearly owes as much to our sky and our detectors as it does to the galaxy itself. We have already noted that the apparent diameter of a galaxy depends on its surface brightness, having a maximum for some particular value of Σ , and the same is also true of the apparent luminosity⁴. The observed quantities of a galaxy refer only to those parts of a galaxy that are visible to detectors and although these 'isophotal' values will be related to the true values, the relationship could be somewhat distant. (Also, the brightest cores will be 'burned out' in photographs because of the limited dynamic range of emulsions).

We can assume that any sample of galaxies is defined by apparent luminosity and size limits $I_{\text{iso}} \geq I_{\text{lim}}$, $r_{\text{iso}} \leq r_{\text{lim}}$, a galaxy having to satisfy simultaneously both criteria to be included. As we move away from the optimum surface brightness, I_{iso} and r_{iso} decrease (for a given total luminosity), so the maximum distance at which the galaxy can get into the sample will also decrease and its 'visibility' (the corresponding volume of space) will decrease rapidly. Depending on the specific sample limits the probability of inclusion of such galaxies in the sample may be a very sharply

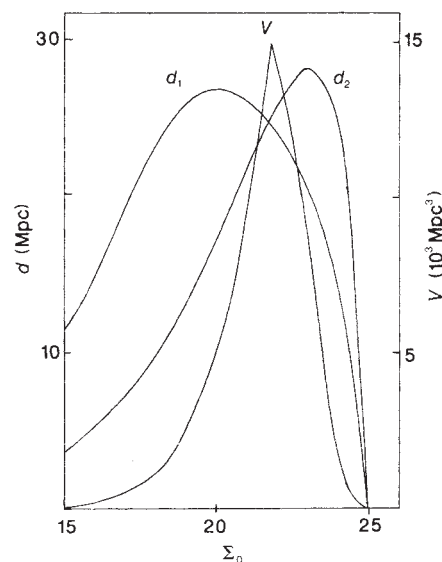


Fig. 2 Calculated visibility of exponential disk galaxies, with a detector sensitive down to 25 magnitudes arc s⁻² in the blue, given as a function of the central surface brightness Σ_0 . d_1 and d_2 are the maximum distances at which galaxies of a particular total luminosity exceed given isophotal luminosity and diameter units, respectively. The visibility, V , is proportional to $[\min(d_1, d_2)]^3$. Note that at the intersection of d_1 and d_2 , the one is rising, the other falling, which explains the sharp peaking of V .

peaked function of surface brightness^{3,4} (for example, see Fig. 2).

The similarity of Figs 1 and 2 suggests that the disk galaxies in well-known catalogues may be only a subset of the whole population. Malin 1 would not get into the RC2 catalogue and but for the presence of its nuclear bulge, this massive disk would have remained undetected whatever its distance, as its central surface brightness is below the limit of all contemporary catalogues. Even in a deep survey with a sensitivity of about 1 per cent of the sky background, galaxies like Malin 1 would be 150 times less likely to appear than disks of similar luminosity but with the canonical central-surface brightness of 21.5 magnitudes arc s⁻².

The unexpectedly sharp peak of the visibility function in Fig. 2 is not specific to just one particular catalogue nor to disk galaxies alone. The discovery of many disks with surface brightness as low as Malin 1 will be no surprise to those aware of just how dramatic the peak can be. A significantly larger galaxy population than is generally assumed would have implications, and could suggest methods of searching for the postulated icebergs.

For example, the number of galaxies along the line-of-sight to quasi-stellar objects (QSOs) will exceed present expectations. This could explain the large number of absorption-line systems, normally attributed to extremely large galactic haloes, that are seen in QSO spectra. Alternatively, one might expect to find radio emission from H I clouds, at a wavelength of 21 cm, that lack obvious

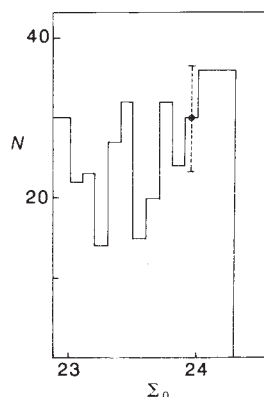


Fig. 3 Distribution of central surface brightness for low-surface-brightness galaxies in the Fornax cluster, after correction for selection effects (data from ref. 10).

optical counterparts. With rare exceptions, searches of such clouds have proved unsuccessful. Malin 1 has been observed at 21 cm, but its hydrogen-mass-to-light ratio is exceptionally high¹ and may not be common to all such galaxies. Similarly, unidentified extended X-ray sources⁷ may be low optical surface brightness objects.

Direct observations of nearby galaxy clusters show up many low-surface-brightness dwarf galaxies^{8,9}. Indeed, down to the limits we have currently surveyed¹⁰ there is no sign of a cut-off, or even a decrease in numbers as we move to very low surface brightnesses, once the selection effects are taken into account¹¹ (Fig. 3). Deeper searches being carried out by Bothun and collaborators⁹ and ourselves may therefore be expected to turn up even fainter objects in large numbers.

The number (and masses) of galaxies in different volumes of space obviously have implications for many current problems such as missing mass, galaxy formation, clustering, biasing and so forth. In view of the sharply peaked form of the visibility function and the way it affects the better-known catalogues of galaxies, no one should be confident that all varieties of galaxy, even in the more accessible regions of space, have yet been discovered. The crouching giant, Malin 1, may be the first of many. □

1. Bothun, G.D., Impey, C.D., Malin, D.F. & Mould, J.R. *Steward Observatory Preprint No. 728* (1987).
2. Malin, D.F. *Am. Astron. Soc. Photo-Bull.* **27**, 4 (1981).
3. Disney, M.J. & Phillips, S. in *Proc. Vatican Conf.* (in the press).
4. Disney, M.J. & Phillips, S. *Mon. Not. R. astr. Soc.* **205**, 1253 (1983).
5. Jura, M. *Astrophys. J.* **238**, 499 (1980).
6. Fish, R.A. *Astrophys. J.* **139**, 284 (1964).
7. Bechtold, J. *et al. Astrophys. J.* **265**, 26 (1983).
8. Sandage, A., Binggeli, B. & Tammann, G.A. *Astron. J.* **90**, 1759 (1985).
9. Bothun, G. in *Proc. Santa Cruz Workshop on Nearly Normal Galaxies* (in the press).
10. Cawson, M.G.M., Kibblewhite, E.J., Disney, M.J. & Phillips S., *Mon. Not. R. astr. Soc.* **224**, 557 (1987).
11. Phillips, S., Disney, M.J., Kibblewhite, E.J. & Cawson, M.G.M. *Mon. Not. R. astr. Soc.* (in the press).

Michael Disney and Steven Phillips are in the Department of Applied Mathematics and Astronomy, University College, PO Box 78, Cardiff CF1 1XL, UK.

Molecular neurobiology

A new type of ion-channel block

David Colquhoun

ION channels are protein molecules that span cell membranes. They can exist in open or shut conformations; when open, ions flow through the channel to produce a current, commonly a few picoamperes in magnitude, across the membrane. Many drugs, such as local anaesthetics, ganglion blockers and anti-dysrhythmics, many heuristic agents, and some ions can block channels and so transiently interrupt the flow of current through them. In all cases reported so far it seems that the blocking molecule causes complete occlusion of the channel, but elsewhere in this issue (*Nature* **329**, 243; 1987), Prod'hom *et al.* report that an unusually simple blocking agent, a single proton, can block calcium channels and that when the proton is bound to the channel, the flow of current through it is reduced but not abolished.

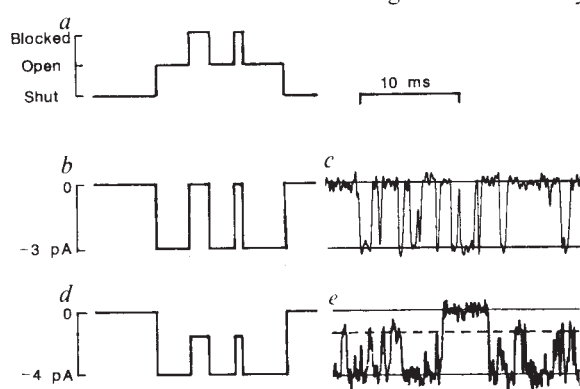
Some kinds of ion channels, receptor-operated channels, open as a result of binding of neurotransmitter molecules to sites on their extracellular aspect. Others, voltage-operated channels, open as a result of a change in the potential gradient across the membrane. Some types of channels, when open, will allow the passage of most small cations; others show a high degree of specificity for one sort of ion. Prod'hom *et al.* studied one type of voltage-operated calcium channel in guinea-pig heart cells. Under physiological conditions, this channel is highly selective for calcium ions, and it is thought that this selectivity results from the presence of two sites in the channel with rather high affinity for calcium ions (Almers, W. & McCleskey, E.W. *J. Physiol., Lond.* **353**, 585; 1984, and Hess, P. & Tsien, R.W. *Nature* **309**, 453; 1984). In the absence of calcium the channel becomes non-selective and many monovalent and divalent

cations can pass through it. Prod'hom *et al.* obtained good resolution in their experiments by using sodium as the charge carrier in the absence of calcium, and in addition, by inclusion of a so-called 'calcium agonist', a drug that lengthens the mean time that the channel is open.

Most sorts of channel are susceptible to blockage and the characteristics of block of different sorts of ion channels often have common features. The current flow through the channel, for example, is usually blocked completely (see figure) during the period when the blocking molecule is bound, and the rates of binding and dissociation of the blocking molecules are usually strongly dependent on the membrane potential (when the blocking molecule itself is charged). The magnitude of this voltage dependence suggests that the blocking molecule binds about half-way through the membrane field so the binding site is presumably deep within the channel; it is probable that the channel is simply plugged by the blocking molecule, a view reinforced by the fact that the blocker can, in some cases, go right through the channel. In many cases it is also found that the ion channel can be blocked only, or at least preferentially, when it is open. This characteristic gives rise to 'use-dependence' of blocking; the more the channel is used (for example, the higher the frequency of action potentials in a neuron), the more often the channel is open and therefore susceptible to becoming blocked. This phenomenon may be of therapeutic value — for example, the high activity in an epileptic focus may be more easily suppressed than normal activity.

Recordings of single ion-channel currents show that blockages occur randomly.

a, Schematic illustration of random transitions between shut state(s), the open state and the blocked state; *b*, the current that would be observed for the transitions shown in *a* when blocker completely occludes the channel; *c*, experimental record from frog endplate nicotinic channels in the presence of decamethonium (50 μ M) which blocks the channel as well as opening it. Many of the interruptions of the current will be the result of blockages (mean duration 2.8 ms) but all shut periods that are long enough to be resolved reach the baseline level (C.G. Marshall, D.C. Ogden & D.C., unpublished results). *d*, The current that would be observed for the transitions shown in *a* when the occlusion of the channel is incomplete. Blockages can be clearly distinguished from other sorts of shut period. *e*, An example of such blockages from the work of Prod'hom *et al.* reported in this issue; blockages of the calcium channel produced by deuterium ions ($pD = 8.2$, or 6.3 nM).



When the blockage is total (b , c in the figure), the current is zero during a blockage just as it is when the channel is in one of its normal shut states, so it is not possible to say unambiguously whether a shut period is a blockage or not. In many cases, however, the blockages are identifiable as a characteristic component in the distribution of the durations of shut times, and from the dependence of their frequency on the free concentration of the blocking molecules. The work of Prod'homme *et al.* is unusual in that the simplest possible chemical agent, a single proton, produces ion-channel block, and that this block is incomplete; when the proton is bound the current is reduced to about one-third of its normal value, so blocked periods are unambiguously identifiable (unless they are very brief) by direct inspection of the record. Another unusual feature of the proton block is that it does not depend on membrane potential; the binding site for the proton must be outside the electric field of the membrane, on the extracellular surface of the channel. The mechanism seems to be quite different from that of most other blockers; rather than the blocker entering some way into the channel and plugging it, the authors suggest that the change of local surface potential produced by protonation reduces the cation concentration near the mouth of the channel and so reduces the flow of ions through the channel.

The pK of 7.4 for the protonation reaction corresponds to an equilibrium constant of about 40 nM for proton binding.

This rather high affinity is similar to that found for block of nicotinic channels by tubocurarine molecules (at -100 mV). The association rate constant for tubocurarine binding to the blocking site has a conventional value, of the order of 10^7 $M^{-1} s^{-1}$, and the mean duration of the blockages (the reciprocal of the dissociation rate constant) is of the order of 2 s (Colquhoun, D. *et al.* *J. Physiol., Lond.* **293**, 247; 1979). Protons, however, are well known to be highly mobile, and the association rate constant for channel block by protons found by Prod'homme *et al.*, 4×10^{11} $M^{-1} s^{-1}$, is four orders of magnitude higher than for most other blockers. It is even slightly greater than that for association of H^+ and OH^- ions (see Eigen, M. & De Maeyer, L. *Proc. R. Soc. A* **247**, 505; 1958). This means that, despite the high equilibrium affinity for protons, they are bound for surprisingly short times, on average about 60 μs (and rather longer, about 150 μs , for deuterium ions). More generally, any proton-binding reaction with a pK value in the range 6–8 (equilibrium constant 1 μM –10 nM), for which effects might be obvious at physiological pH, is likely to produce brief bindings (2.5–250 μs) despite the rather high affinity. It will be interesting to see whether effects like this contribute to the extra noise that appears when many sorts of channel open, or to the appearance of subconductance levels. $\square$

David Colquhoun is in the Department of Pharmacology, University College London, Gower Street, London WC1E 6BT, UK.

Plant ecology

Moving in on the gaps

Peter D. Moore

NATURE abhors a vacuum, but the creation and refilling of vacancies in plant communities is important in the maintenance of species diversity in various habitats. Recent studies ranging from tropical forests¹, temperate woods², pine barrens³ and assemblages of colonizing annuals⁴ all confirm that the invasion of gaps in vegetation follows predictable patterns, depending on the requirements of the invasive species, and provides opportunities in both space and time for many early successional plants to maintain viable, if scattered, populations.

The idea of the 'regeneration niche' and its contribution to the richness of plant communities has been set out most clearly by Grubb⁵. The concept recognizes the dynamic nature of vegetation in which the death of mature individuals, particularly of large-bodied dominants such as trees in a forest, provides a new, if temporary, environment in which species with shorter life-spans and generation times take

advantage of the more open conditions. The result is that many species associated with early stages in a succession can maintain themselves in a 'climax' system.

Canopy gaps undoubtedly contribute to the diversity of tropical forests, where gap turnover rates have been estimated at around 70–160 years (ref. 6). Studies such as those of Brokaw¹ on Barro Colorado Island in Panama confirm the existence and variety of regeneration niches in this setting. Invasive tree species following the creation of a gap vary in their time of appearance and in their gap-size requirements. *Trema micrantha* made its appearance within a year of gap formation, grew very rapidly (up to 7 metres per year), but failed to survive beyond the ninth year in any but the largest gaps. *Cecropia insignis* also arrived early, within the first three years, but grew slowly and occurred only in the larger gaps (more than 215 metres). In the same area, *Miconia argentea* could be recruited into

the gap community up to the seventh year independent of gap size. Niche overlap is thus apparent, but the varied preferences of invasive species contribute to overall forest diversity.

Any contribution to the diversity of plant species will inevitably provide new opportunities for animals, for example, the production of succulent, edible fruits by many of the gap-regeneration species. This question has recently been reviewed by Leck⁷, who finds that there are more bird-dispersed fruit trees in younger forests than in older ones, and that bird communities of these second-growth areas are indeed richer as a consequence. Of particular interest is the importance of such secondary growth in supporting migrant flocks of fruit-eating birds retreating from winter conditions in temperate North America.

It is not just in the tropical forests that migratory frugivores enrich the avifauna of the gaps. Skeate⁸ has examined the relationships between birds and fruit in the so-called hammock communities of hardwood stands in the prairie, marsh and pine forests of Florida. Not only are gap species important in providing food for migrating birds, but their sequence of flowering is closely synchronized to the migratory pattern of their consumers. The gap relationships of the fruit-producing plants concerned, such as *Cornus florida* and *Aralia spinosa*, have been examined by Runkle and Yetter², who show that the former species is most dominant in larger gaps and the latter in older gaps. The regeneration patches in woodland provide a sequence of food-rich habitats for migrating birds along their entire routes to the tropics.

The importance of gaps and their regeneration as contributors to the animal and plant diversity of forests is thus evident from this type of work. What are the mechanisms of invasion and facilitation in the regenerative process? Are the bird-dispersed shrubs dependent on their avian vectors for dispersal into gaps? Or, like the pin cherry (*Prunus pensylvanica*), a gap species of the temperate woodlands, are they able to maintain seed banks over the 70 or more years between germination opportunities? Information is clearly needed about the respective roles of coevolution and of seed dormancy in the spatial and temporal complexity of the canopy gap. $\square$

1. Brokaw, N.V.L. *J. Ecol.* **75**, 9–19 (1987).
2. Runkle, J.R. & Yetter, T.C. *Ecology* **68**, 417–424 (1987).
3. Gibson, D.J. & Good, R.E. *Oikos* **49**, 91–100 (1987).
4. MacConnaughay, K.D.M. & Bazzaz, F.A. *Ecology* **68**, 411–416 (1987).
5. Grubb, P.J. *Biol. Rev.* **52**, 107–145 (1977).
6. Clark, D.A. *Trends Ecol. Evol.* **1**, 150–154 (1986).
7. Leck, C.F. *Trends Ecol. Evol.* **2**, 33 (1987).
8. Skeate, S.T. *Ecology* **68**, 297–309 (1987).

Peter D. Moore is Reader in Ecology, Department of Biology, King's College London (KQC), Campden Hill Road, London W8 7AH, UK.

Geochemistry

Mid-ocean-ridge basalt genesis

R. N. Thompson

WHAT do the topographic depths of mid-ocean ridges, the thickness of their crusts and the compositions of their extruded lavas have in common? A great pleasure of science is the way simple unifying principles emerge to link a series of seemingly disparate parameters into a coherent pattern; even if the simplicity of the first attempt becomes less apparent after future research. It is agreed that the Earth's silicate mantle convects and that — at least beneath mid-ocean ridges — this leads inevitably to partial melting of peridotite that upwells within a few tens of kilometres of the surface, to yield liquids that proceed buoyantly to eruption or near-surface intrusion. The amount and composition of this melt varies according to the temperature of the upwelling peridotite (if its composition is essentially constant). Lateral variation of the mantle heat beneath mid-ocean ridges also controls their topographic depths. Thus, the three parameters listed above could all be related to one underlying principle. In their recent work¹, Klein and Langmuir are remarkably successful in analysing the depths and crustal thicknesses of mid-ocean ridges, and the compositions of their lavas, along these lines.

Klein and Langmuir start with major-element analyses of basalt glasses from ocean ridges throughout the world. To compare the compositions of these, they first correct for the effects of the variable fractional crystallization of each magma since its genesis. To do this they recalculate each composition to a fixed MgO content (approximately equivalent to a single magmatic temperature), using a simple empirical closed-system fractional crystallization model. Those who are convinced that the compositions of the individual upwelling magma batches beneath ocean ridges are indecipherably scrambled, by open-system multiple magma-mixing and crystallization events in near-surface reservoirs², will probably criticize this first stage. Nevertheless, Klein and Langmuir find that their simple normalization produces excellent world-wide correlations between water depth along the mid-ocean ridges and such parameters as Na₂O and FeO content, and the ratio CaO/Al₂O₃.

Melting experiments

Klein and Langmuir compare these trends with those observed within the liquids produced in experiments on the polybaric fusion behaviour of natural peridotites. This stage of the argument is complicated because a single parameter

can often reflect more than one process. Thus Na₂O content is sensitive to both the extent and the pressure of fusion during partial melting of peridotite. But FeO in the liquid is remarkably insensitive to the degree of isobaric melting in mantle peridotite (although sensitive to the pressure), and this enables Klein and Langmuir to decide which of several explanations of their composition–depth trends is correct.

It emerges that basalts from shallower-than-average segments of the global mid-ocean ridge system, such as the Kolbeinsey Ridge, are distinguished by relatively low Na₂O, SiO₂ and Ni contents and ratio Sm/Yb, and high FeO, Sc and CaO/Al₂O₃. This seems to be because they are generated by above-average fusion of the subaxial mantle, with the melt escaping from its source at greater-than-average mean pressures. Basalts from deeper-than-average segments of the ridges, such as the mid-Cayman Rise, show contrasting chemical features attributed to below-average mantle fusion and shallower-than-average melt escape. Between these two extremes there is a remarkably systematic variation between basalt chemistry and ridge depth.

Links

What do these links between ridge depth, degree of mantle fusion and mean depth of magma escape reveal about the convective behaviour of the upper mantle? As Klein and Langmuir point out, all these parameters are related to lateral variations in the heat of upwelling mantle beneath spreading centres. Thus, locally hotter mantle — in a plume or jet — begins to melt at a greater depth during adiabatic uprising, and eventually melts to a larger extent, than the relatively cooler mantle between jets. Both the locally greater crustal thickness (formed by melt products) and the direct effect of a hotter-than-average upwelling jet contribute to the shallowness of the sea above such regions.

The approach used by Klein and Langmuir — analysing the compositional variations of ocean-ridge basalts — is one way to quantify the melt output from upwelling mantle. The other approach uses the experimental phase-equilibria data obtained for peridotites³. These provide the compositions of isobaric, polythermal liquids which can be re-cast to give the polybaric liquids that form within a volume of partially molten mantle peridotite undergoing isentropic upwelling⁴. The upper limit of mantle fusion by adiabatic

decompression along mid-ocean ridges is generally agreed to be 20–25 per cent melting, perhaps slightly more for different bulk compositions atop the convective jets that underlie such places as the Azores and Galapagos^{1,3,4}. Between the hottest and coldest parts of the upwelling mantle there seems to be a potential (pressure-corrected) temperature difference of only 200–250°C (refs 1, 4).

Applications

There are perhaps two main applications of these studies and two important areas for future work. As Klein and Langmuir point out, the compositions of lavas from old parts of existing ocean basins could be used to deduce the former distribution of convective jets in the mantle beneath oceans. McKenzie and Bickle⁴ quantify the effects of the stretching of pre-existing lithosphere, so that the amount and composition of the consequent magma may be deduced from knowledge of the thickness of lithosphere, the amount of its extension and the potential temperature of the subjacent convecting mantle. Further studies may clarify two questions. First, as the partially molten mantle upwells, does all the released melt percolate upwards through the deforming matrix (mixing with other melts produced nearer the surface) and emerge as a 'magmatic cocktail' at the top of the melting column, or can some liquid batches escape from the mantle residuum at depth and proceed to the surface in a state of physical isolation or near isolation?

Second, how can the behaviour of strongly incompatible elements — those which have very small bulk-distribution coefficients between melt and residuum during mantle fusion — be quantified? Klein and Langmuir and many others^{4–8} have deduced that these elements obey a different set of rules from the major and compatible trace elements during fusion of upwelling mantle. But we are not yet able to quantify the rules. To do so would help in clarifying the geochemistry of the three main radiogenic isotope systems — Sr, Nd and Pb — upon which most of our understanding of the time-integrated behaviour of the convecting mantle depends. All their parent and daughter elements are strongly incompatible in mantle-melt systems. □

1. Klein, E.M. & Langmuir, C.H. *J. geophys. Res.* **92**, 8089–8115 (1987).
2. O'Hara, M.J. & Mathews, R.E. *J. geol. Soc. Lond.* **138**, 237–277 (1981).
3. Thompson, R.N. *Earth Sci. Rev.* (in the press).
4. McKenzie, D. & Bickle, M.J. *J. Petrol.* (in the press).
5. Thompson, R.N., Morrison, M.A., Hendry, G.L. & Parry, S.J. *Phil. Trans. R. Soc. A310*, 549–590 (1984).
6. O'Hara, M.J. *Nature* **314**, 58–62 (1985).
7. McKenzie, D. *Earth planet. Sci. Lett.* **72**, 149–157 (1985).
8. Galer, S.J.G. & O'Nions, R.K. *Chem. Geol.* **56**, 45–61 (1986).

R.N. Thompson is Professor of Geology at the University of Durham, South Road, Durham DH1 3LE, UK.

HBLV (or HHV-6) in human cell lines

SIR—Last year we described the isolation from patients with lymphoproliferative disorders of a novel human herpesvirus, human B lymphotropic virus (HBLV), which infects freshly isolated human mononuclear cells from cord blood, adult peripheral blood and spleen but not from a number of established human and animal cell lines^{1,2}. The requirements for repeated infection of fresh lymphocytes as the target population presented a hindrance to biological, immunological and molecular studies. Now that we have found several cell lines that can be productively infected with the virus we would like to describe them for the sake of those who might find them of value for research.

Five such cell lines are listed in the table. All of them can be infected by HBLV with either cell-free virus or by cocultivation with previously infected cord blood cells. HSB-2, a T-cell line, was highly susceptible to infection by HBLV. By nine days after infection > 90% of the cells are fatally infected. Morphologically, the cells resemble HBLV-infected cord blood mononuclear cells¹ and diffuse staining of the entire cell is obtained in indirect immunofluorescence assay (IFA). The amount of HBLV released by day 9 into supernatant fluids of infected HSB-2 cells is 10^3 – 10^4 TCID₅₀/ml⁻¹ assayed on fresh cord blood mononuclear cells and HSB-2 cells.

The megakaryocyte cell line, HEL, showed a peak of IFA-positive cells (~30 %) by 9–10 days after infection, with predominantly nuclear fluorescence. HEL cells did not demonstrate a clear morphologic change following infection. Furthermore, only a low per cent of HEL cells continued to express HBLV proteins and release virus. By day 45, the number of IFA-positive cells was less than 5 %. It is possible that these residual of HEL cells, like some EBV cell lines, are persistently but nonproductively infected.

The other three cell lines shown in the Table are substantially less susceptible to HBLV infection. In two of them, the infection persists at a low level for several months, with a continuing release of low

levels of infectious HBLV but virus becomes undetectable in HTB-14 cells within 14 days, suggesting lytic infection.

Cell clones were established from the WIL-2 cell line by limiting dilution and tested for susceptibility to HBLV infection. In a selected cloned cell population (ET62), > 90 % of cells were infected and released high levels of infectious virus.

High molecular weight DNA extracted from the HBLV-infected HSB-2, HEL, ET62 and HBT-14 cells yielded characteristic restriction endonuclease banding patterns by Southern blot analysis, whereas uninfected cells were negative.

Of the five cell lines, HSB-2 and HEL should be the most useful. They are free of other herpesviruses and are therefore more suitable for immunovirologic and molecular analysis and for developing HBLV-specific reagents. Moreover, HSB-2 should be an excellent source for large-scale production of HBLV. Such reagents may lead to a linkage of HBLV with one or more human diseases. Our virus-producing cell lines will be made available to investigators interested in research on HBLV. We feel that the virus should be classified HHV-6 (human herpes virus-6) in accord with the published provisional classification of herpesviruses⁷.

D.V. ABLASHI
S.Z. SALAHUDDIN
S.F. JOSEPHS
F. IMAM
P. LUSSO
R.C. GALLO

National Cancer Institute,
National Institutes of Health,
Bethesda, Maryland 20892, USA

C. HUNG
J. LEMP

Electro-Nucleonics, Inc.,
Silver Spring, Maryland 20904, USA

P.D. MARKHAM

Bionetics Research, Inc.,
Kensington, Maryland 20850, USA

1. Salahuddin, S.Z. *et al. Science* **234**, 596–601 (1986).
2. Josephs, S.F. *et al. Science* **234**, 601–603 (1986).
3. Ponten *et al. Acta path. Microbiol. Scand.* **74**, 465–486 (1969).
4. Foley *et al., M.D. Anderson Hospital and Tumor Inst. Monograph* **21**, (1967).
5. Levy, J.A. *et al. Cancer* **22**, 517–524 (1968).
6. Buell, D.N. *Ann. N. Y. Acad. Sci.* **190**, 221–234 (1972).
7. Roizman, B. *IRA IRAC Monographs* **19** (1979).

Table 1 Infection of continuous human cell lines with HBLV

Source	Cell line identification	HBLV infectivity* titre (TCID ₅₀ /ml)	Average number of HBLV infected cells (IFA)(%)	Endogenous herpesviruses
1	HSB-2	10^3 – 10^4	90 (day 9)	None
2	HEL	10^1	30 (day 10) 5 (persistent)	None
3	HTB-14	None	3 (day 7)	None
4	WIL-2	~ 10^{1-2}	2–5 (persistent)	EBV
5	IM-9	~ 10^{1-2}	2–5 (persistent)	EBV

*Infectivity titres were performed on human cord blood mononuclear cells according to the procedure described previously¹. Source: 1, T-lymphoblastoid cell line⁴ (ATCC CCRF-HSB-2); 2, megakaryocytic cell line from normal human bone marrow (gift from Doris Morgan, Hahnemann University, Philadelphia); 3, glioblastoma cells³ (ATCC); 4, 5 HTB-14 B-lymphoblastoid cell line⁶ (ATCC).

Comments on the sombre view of AIDS

SIR—We want to comment on the method Rees¹ used to estimate the mean incubation period of transfusion-associated AIDS (acquired immune deficiency syndrome).

First, Rees did not consider the left censoring of transfusion-associated AIDS data. Because the association between AIDS and blood transfusions was not recognized until the middle of 1982^{2,3}, cases of AIDS with onsets before 1982 may have been missed. Consequently, Rees might have overestimated the true mean. (Rees's Table 1 contains cases with diagnosis dates between 1981 and 1982, because he transformed the crude data reported by Peterman⁴.)

Second, Rees considered only the densities in the family of normal distributions that have a ratio of mean to standard deviation equal to 3. If this restriction on the ratio is removed, then other normal distributions could be found that have only a minor increase of standard deviation, means much shorter than 15 years, and total absolute differences smaller than 14.67 (the smallest total absolute difference presented in Rees's Table 3).

Third, Rees focused only on the fit of the assumed model to the marginal observed frequencies, and there was no interval estimate or standard error of the estimate. Furthermore, he stated no statistical optimal properties for the method.

Finally, another approach is to use the maximum likelihood estimator based on the truncated Weibull model². This method can take into account the left and right censoring. Using this approach, we calculated an estimate of 4.5 years for the mean incubation period with a 90% confidence interval of 2.6 to 14.2 years. A recent calculation based on current data suggests that the mean incubation period may be even longer than 4.5 years. However, with a limited observation time and a very long mean incubation period for transfusion-associated AIDS, it is very difficult to pinpoint the mean at present.

KUNG-JONG-LIU

Division of Injury Epidemiology and
Control,

Center for Environmental Health,

THOMAS A. PETERMAN

AIDS Program,

DALE N. LAWRENCE

Division of Host Factors,

Center for Infectious Diseases,

Centers for Disease Control,

Public Health Services,

US Department of Health and Human

Services,

Atlanta, Georgia 30333, USA

1. Rees, M. *Nature* **326**, 343–345 (1987).
2. Lui, K.J. *et al. Proc. natn. Acad. Sci.* **83**, 3051–3055 (1986).
3. Curran, J.W. *et al. New Engl. J. Med.* **310**, 69–75 (1984).
4. Peterman, T.A. *et al. J. Am. med. Ass.* **254**, 2913–2917 (1985).

Cometary organics but no evidence for bacteria

SIR—Hoyle and N. Wickramasinghe¹ compare the 3–4 μm spectrum of Comet Halley obtained at the Anglo-American telescope by D. Wickramasinghe and Allen² with the predictions of a bacterial model. We believe that the agreement between the two provides no evidence for the existence of cometary bacteria.

A variety of plausible cometary materials exhibit the observed 3.4 μm emission feature. These include the organic residue of irradiated candidate cometary ices³ (such as the residue of irradiated methane ice clathrate) and polycyclic aromatic hydrocarbons⁴, of which the residue may be partly composed. Indeed, any molecule containing $-\text{CH}_3$ and $-\text{CH}_2-$ alkanes will emit at 3.4 μm under suitable conditions. Therefore tentative identifications must rest on additional evidence, including a plausible account of the origins of the organic material, a plausible model for the infrared emission of this material, and a demonstration that this conjunction of material and model not only matches the 3–4 μm spectrum, but also does not yield additional emission features where none is observed. In the case of the residue of irradiated low-occupancy methane ice clathrate, we have argued⁵ that the synthesis of the organic residue in the laboratory well simulates the radiation processing experienced by Comet Halley, and have shown that a simple model based on the spacecraft-determined dust distribution in the Halley coma fits the 3.4 μm feature, provides optical depths in excellent agreement with those observationally determined and accounts for the absence of features at longer wavelengths. Thus any attempt to explain the spectrum of Comet Halley with living organisms or their products seems an extravagant departure from Occam's Razor.

Hoyle and Wickramasinghe neither explain nor reference their modelling procedure. The final sentence of their correspondence leaves the impression that such an account will be found in their ref. 3, but in fact this is the paper giving the observed Halley spectrum², which certainly reports no bacterial modelling. It is thus difficult to evaluate critically their results (shown graphically). Nevertheless a number of difficulties present themselves.

Hoyle and Wickramasinghe match the Halley data with the emission spectrum of bacteria at 320 K. However, the continuum in the 3–4 μm spectrum is best fitted² by summing the observed flux from scattered sunlight with a black body at 350 ± 10 K; 350 K, adjusted for heliocentric distance, is also the continuum temperature found by Danks *et al.*⁶ in the 3–4 μm range, as well as (to within the 10 K uncer-

tainty) that which best fits the 5–13 μm continuum⁷. Thus the temperature used by Hoyle and Wickramasinghe lies outside the observational error bars. Moreover, the feature at ~ 3 μm , which is somehow well-fitted by the bacterial emission spectrum, is revealed by both ground-based observations² and those made by the Vega spacecraft⁸ to be an absorption feature (probably the O–H vibrational transition) lying well below the continuum flux.

The choice of a single temperature to characterize both the continuum and the emission feature, as well as the description of their model curve as simply "the predicted spectrum of bacteria heated to 320 K", indicates that Hoyle and Wickramasinghe are in effect treating the entire 3–4 μm spectrum (both emission feature and continuum) as due to emission from bacteria. However, there are grave objections to such a procedure. The intensity of the 3.4 μm band varies independently of the continuum by as much as a factor of 4.5, suggesting that the continuum and emission feature are produced by two independent dust components² (indeed, this is also the interpretation given to the two-fold variation of the 10 μm silicate dust feature relative to the continuum⁷). The continuum flux must be due primarily to the $\sim 98\%$ of the coma dust-filling factor provided by grains ≥ 1.0 μm in radius.⁵ Grains with radii ≥ 10 μm , comprising $\sim 77\%$ of the coma-filling factor, will emit as black bodies (being individually optically thick at 3–4 μm); the bacterial spectrum is in any case irrelevant for these. The 3.4 μm emission feature, on the other hand, must be due to grains that are individually optically thin at this wavelength, that is the $\sim 2\%$ of the dust-filling factor due to grains with radii ≤ 1.0 μm . Hanner⁹ has demonstrated that such submicron cometary dust will have temperatures hundreds of K hotter than a black body at 1 AU. Thus an appropriate temperature for the submicron emitting grains is probably ~ 500 K, not 320 K.

Both spacecraft and ground-based near-infrared observations, as well as the results of *in situ* mass spectroscopy¹⁰, strongly suggest the presence of complex organic or organic-coated grains in the Halley coma. It would be unfortunate if claims for living organisms prejudiced the reception of these important results.

CHRISTOPHER CHYBA
CARL SAGAN

Laboratory for Planetary Studies,
Cornell University,
Ithaca, New York 14853, USA

1. Hoyle, F. & Wickramasinghe, N. C. *Nature* **328**, 117 (1987).
2. Wickramasinghe, D. T. & Allen, D. A. *Nature* **323**, 44–46 (1986).
3. Khare, B. N. *et al.* *Icarus* (in the press).
4. Knacke, R. F. *et al.* in *20th Esab Symposium on the Exploration of Halley's Comet*, 95–100 (ESA SP-250, 1986).
5. Chyba, C. & Sagan, C. *Nature* (in the press).
6. Danks, A. *et al.* in *20th Esab Symposium on the Exploration of Halley's Comet*, 103–106 (ESA SP-250, 1986).
7. Bregman, J. D. *et al.* *Astr. Astrophys.* (in the press).
8. Combes, M. *et al.* *Adv. Space Sci.* (in the press).
9. Hanner, M. S. in *Cometary Exploration II* (ed. Gombosi, T. I.) 1–22 (Hungarian Acad. Sci., 1982).
10. Kissel, J. & Krueger, F. R. *Nature* **326**, 755–760 (1987).

The blood–brain barrier

SIR—The endothelium of cerebral blood capillaries forms a cellular barrier with selective properties that isolates circulating compounds in the blood from those produced by the brain, so maintaining the homeostatic environment of the brain^{1,2}. The perivascular glial cells may play a role in the induction and maintenance of these barrier properties.

In elegant experiments, Janzer and Raff³ recently provided clear evidence for the primary role of type I astrocytes in the induction of impermeability to macromolecules in the capillaries of the iris. The finding raises several intriguing questions including the nature of the inducing signal and the importance of cell-to-cell communication. I think, however, that there is an important reservation about the interpretation of these experiments, as Brightman and Reese⁴ and Reese and Karnovsky⁵ have shown that the cerebral endothelium differs in structure from that of peripheral vessels in two ways: first, in cerebral vessels there are highly impermeable tight junctions between endothelial cells, and second, the cells have much lower pinocytotic activity than peripheral endothelial cells. It is these two features that give the cerebral endothelium its selective properties. Vasoactive substances, such as histamine can activate pinocytosis⁶ without opening the tight junctions⁷ indicating that activation of pinocytosis alone is sufficient to change the permeability of the barrier. The low rate of pinocytosis may also contribute to the impermeability of these vessels to macromolecules⁸.

The model of Janzer and Raff will certainly be useful for further studies of the transformation of a peripheral endothelium into one with features of cerebral capillaries. I consider, however, that fine structural investigations are essential to determine whether the endothelial cells have cerebral or peripheral characteristics. Without this information it is not possible to tell with certainty which of the structural features determines development of impermeability to macromolecules. But it would not surprise me if the suppression of pinocytosis turns out to be the worthy side of the coin.

FERENC JOÓ

National Institute of Neurological and
Communicative Disorders and Stroke,
National Institutes of Health, Bethesda,
Maryland 20892, USA

1. Bradbury, M. *Concept of a Blood–Brain Barrier* (Wiley, Chichester, 1979).
2. Cserr, H. F. & Bundgaard, M. *Am. J. Physiol.* **246**, R277 (1984).
3. Janzer, R. C. & Raff, M. C. *Nature*, **325**, 253 (1987).
4. Brightman, M. W. & Reese, T. S. *J. Cell Biol.* **40**, 648 (1969).
5. Reese, T. J. & Karnovsky, M. J. *J. Cell Biol.* **34**, 207 (1967).
6. Dux, E. & Joó, F. *Expl Brain Res.* **47**, 252 (1982).
7. Dux, E. *et al.* *Neuroscience*, **12**, 951 (1984).
8. Joó, F. *Nature* **321**, 197 (1986).

New world of knowledge

I. Bernard Cohen

The Launching of Modern American Science 1846–1876. By Robert V. Bruce.
Knopf: 1987. Pp.446. \$30.

ROBERT V. BRUCE, emeritus professor of American history at Boston University and author of a biography of Alexander Graham Bell, has turned his attention to the formation of the American scientific community between 1846 and 1876. The year 1846 is associated with the Smithsonian Institution (America's first national institute of science, established through a private bequest), America's first scientific school (at Yale; Harvard followed a year later), and the arrival in America of Louis Agassiz. The year 1876 is notable for the founding of the American Chemical Society (a token of the maturing of an American scientific speciality) and of the Johns Hopkins University (the first in America to be dedicated to the German ideal of 'higher learning' and advanced degrees, coupled with research). The terminal year of this study, 1876, was significant also for the publication of J. Willard Gibbs's great monograph *On the Equilibrium of Heterogeneous Substances*. The three decades surveyed witnessed not only the professionalization of scientific education in America, the beginning of graduate study in the sciences and the beginning of serious financial support for scientific research; they were also the years of the emergence of a sense of community among American scientists and the coming into being of the National Academy of Sciences — national public recognition that scientists might perform a function essential to society.

In an introductory chapter Bruce admits that the "history of the scientific enterprise in nineteenth-century America requires some account of its output", but "measured against what Europeans were doing that output was modest". He himself places emphasis "on the process rather than the product, on the internal sociology, economics, and politics of science and on its interaction with the larger society". The reason: "In these lies the real significance of the period".

The story told by Bruce is that of a shift from an enterprise conducted by amateurs who often had little contact with one another, save in small local or regional groups, to an organized community of professionals with shared goals and communal standards. In 1846 15 per cent of the leading scientists in America (those listed in the *Dictionary of American Biography*) were "simon-pure amateurs, drawing no income at all from science-related work", while "slightly more than

half of all *DAB* scientists in 1846" were teachers. By 1876 the role of the self-supporting amateur was dwindling. Universities were coming to recognize that scientists needed to do research as well as teach. The scientific bureaux of state and federal governments were beginning to provide real opportunities for scientific employment. Finally, there were signs that private philanthropy would support scientific work.

By the time of the Civil War nearly a third of the leading American scientists



Powerful presence — the statue of Joseph Henry at the portals of the Smithsonian; in 1846 Henry was appointed secretary of the newly formed Institution.

were employed by the federal government. The most powerful figures in the American scientific scene — such as Joseph Henry and Alexander Dallas Bache — were members of the federal scientific establishment. In the post-war years, science grew in importance in the established universities. Bruce finds that the introduction of the "elective system" actually fostered science because, while it "decreased the proportion of students taking science courses", "those who did take them by choice (or at least by preference) were more apt to bring out the best in their teachers". Here we must take account not only of older institutions such

as Harvard and Yale, but the newer ones such as Cornell and, notably, the new state colleges and universities which came into being as a result of the Morrill Land Grant College Act of 1862.

In mid-century the only way to gain an introduction to research and to pursue graduate study (the two defining features of the true university) was for the American to study abroad, notably in Germany. Bruce notes:

In 1871 the University of Pennsylvania followed Yale's lead and became the second American school to grant the Ph.D. A year later, Harvard established its "Graduate Department", which in 1873 produced two Ph.D.s and an Sc.D. In 1876 James Dana's son Edward, back from chemical study at Heidelberg and Vienna, took his Ph.D. at Yale after all. By then twenty-five American institutions had awarded forty-four Ph.D.s.

Numbers alone do not ever tell a whole story. Bruce observes that the last number may be inflated by the possibility that some of these PhD degrees "may have been gratuitously bestowed on faculty members to tone up the catalogue". It is important to note, however, that long after the 1870s American scientists continued to find it necessary to do some advanced study abroad. Even after the period covered by Bruce — that is, as late as the 1920s and early 1930s — those American scientists who were foremost in research generally had served some apprenticeship in Europe.

Although the expansion of the federal government during the Civil War increased the number of scientists in Washington, in 1876 the Washington scientific community was still second to that of Boston and Cambridge — to judge by statistics of the American Association for the Advancement of Science. In that year there were 45 AAAS Fellows from the Boston-Cambridge area compared with Washington's 27. But Washington had pulled ahead of New York, with 23 Fellows, New Haven with 17, and Philadelphia with 14. Not until the end of the century "would state universities make the Midwest a major scientific region".

By 1876 American scientists had increased in numbers until approximately 3,000 individuals were using some form of scientific training or expertise in their work. But, as Bruce has counted, only "about 370 of them were distinguished enough to make the *Dictionary of American Biography*". Of these almost half were in the earth sciences or the life sciences, about a third were in chemistry or physics, and the remaining small proportion in mathematics and astronomy. But such figures alone do not tell us very much, because they give no indication of the level of training of such scientists nor their achievement on a world scale.

This is indeed the weakness of Bruce's

Smithsonian Institution

presentation. He has amassed an incredible amount of information on a very large number of individual scientists and their activities. He tells us about the backgrounds, formal training or educational experience, employment, and professional association with other scientists and with the larger community. He traces the formation of scientific institutions and measures their significance and he compares the activities of different regional groups of scientists. But it is in keeping with his stated aim that he generally makes no real evaluation of the scientific contributions of the individual scientists he discusses.

Bruce introduces many of the scientists in his book with brief physical characterizations. James Dwight Dana, the geologist, "was a small, slender man, full of nervous energy . . . and winning in his smile". Ormsby MacKnight was a "small, restless Ohioan". Joseph Leidy was a "heavily built man, slightly round-shouldered, but with a face described by one of his friends as of the 'Christ type' ". Asa Gray was "boyish in his short, slight, agile body and clean-cut beardless face". Charles Henry Davis "was a tall, spare man of military bearing, with a high forehead, and aquiline nose, and kindly eyes". But we are told nothing about Davis's main contribution to science, his scholarly edition in English of Gauss's *Theoria Motus*, which made available that master's method of using least squares. Nor are we much enlightened concerning George W. Hill's real achievement to be told that Henri Poincaré said it "contained the germ of all subsequent progress in celestial mechanics", which does not even contain a hint that Hill had been working on problems of the motion of the Moon. The otherwise uninitiated reader will not really be helped much in understanding either the nature of Henry Rowland's great discovery or its significance by the rather vague (and somewhat misleading) statement that "he had made a discovery of some significance for later electron theory". Can one truly understand the emergence of an American scientific community without some evaluation of the worth of the contributions to knowledge on some comparative scale?

Bruce concludes that by the time of the centennial year, 1876, "all the key elements of the modern American scientific establishment had . . . made at least an initial appearance". The "image of American science" invoked by an examination of the 1870s "is that of a newborn foal, callow, ungainly, yet fully formed and swiftly gaining strength. It still had a long way to go. But it had some definite ideas of how to get there". □

I. Bernard Cohen is the Victor S. Thomas Professor (Emeritus) of the History of Science, Harvard University, Cambridge, Massachusetts 02138, USA.

Bad blood

D.R. Higgs

The Molecular Basis of Blood Diseases.

Edited by George Stamatoyannopoulos, Arthur W. Nienhuis, Philip Leder and Philip W. Majerus. Saunders: 1987. Pp. 747. \$76, £69.

ABOUT 12 years ago, Francis Crick came to lecture at the medical school where I was training. Inevitably, there was a large audience and there were even scuffles at the door when the lecture hall became full and the latecomers were asked to follow the talk on a video screen in an overflow room. The event got off to a bad start when the eminent professor introducing Crick temporarily forgot his name, and at the end of the lecture there was an embarrassing dearth of questions. This early encounter between molecular biology and medicine thus turned out to be a sadly inauspicious occasion.

How rapidly times have changed, and no more so than in the field of haematology. The first impact of molecular biology on medical science came in the late 1970s, when the human globin genes were isolated using complementary DNA synthesized from the naturally purified globin RNA present in red blood cells. Subsequent analysis of the common mutants of globin synthesis (thalassaemias) provided a comprehensive catalogue of the mechanisms that underlie single gene disorders. By the early 1980s various techniques had been developed which allowed other specific sequences to be isolated from complex mixtures of mRNAs, and the lessons learned from the study of globin genes were rapidly applied to other haematological disorders.

Thus during the past, remarkable decade we have witnessed fundamental changes in our understanding of the immune response, haemostasis and the pathogenesis of haematological malignancy. This information is not only intellectually satisfying but is already being exploited for the benefit of patients, both in diagnostic services and through the production of the genetically engineered coagulation factors, fibrinolytic agents, lymphokines and haemopoietic growth factors.

The Molecular Basis of Blood Diseases summarizes and critically evaluates these developments. For those with no background in the subject, an introductory chapter explains the terms and methodologies used. Not surprisingly, about a quarter of the book deals with the globin genes which are still the most comprehensively studied of all eukaryotic systems. Perhaps the most interesting unsolved problem in this field concerns the mechanism by which the expression of the

globin genes is controlled during development. This complex subject is dealt with in an excellent chapter by two of the editors. Another large section of the book summarizes our current knowledge of the immunoglobulin and T-cell receptor genes, and contains chapters on the proposed mechanisms underlying haematological malignancy (a clear, timely account of human T-cell leukaemia viruses among them). A third section covers haemostasis, including chapters on inherited bleeding diatheses and fibrinolysis. In addition to the contributions on these main topics, there are also chapters on haematological problems that are just starting to yield to molecular biology, such as the mechanism of complement action, iron metabolism, phagocytosis and platelet function.

This is a beautifully produced book, and with one or two exceptions the standard of writing and presentation is very high. There will clearly be room for expansion in future editions, however; for example there are no chapters on the red cell membrane, or the red cell enzymes which are frequently involved in genetic blood disorders.

Nevertheless, like the *New York Times*, the current version contains "All the news that's fit to print". Furthermore, the editors have succeeded in their attempts to ensure that the information is up to date by "enlisting the leading investigators in hematology". The fact that only one of the 31 authors is from a British university is a conspicuous reminder of the declining influence of Britain's research community.

Because the application of molecular biology to medicine is such a recent event, it is only to be expected that many practising physicians still regard it with a mixture of awe and suspicion, doubting whether it will have much effect on the management of patients. I believe that in the history of medicine the impact of molecular biology will be compared to the development of the microscope in the seventeenth century, in that it will open up new vistas in all branches of medicine. It will eventually change the way in which we think about disease processes, how we analyse them, classify them, and devise strategies to prevent and treat them (take AIDS, for example). *The Molecular Basis of Blood Diseases* is among the first textbooks to examine the effect of molecular biology on the understanding of disease mechanisms. Although it will appeal primarily to those interested in blood disorders, it should also be consulted by all those who are fascinated by the mechanisms that underlie human genetic diseases. □

D.R. Higgs is Clinical Research Fellow in the Medical Research Council Molecular Haematology Unit, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK.

Insult and injury

Stephen H. Schneider

The Greenhouse Effect, Climatic Change, and Ecosystems. SCOPE 29. Edited by Bert Bolin, Bo R. Doos, Jill Jager and Richard A. Warrick. Wiley: 1987. Pp.541. \$110, £56.

In October 1985, an international conference of potentially major importance took place in Villach, Austria. The first conclusion of the conference report read:

Many important economic and social decisions are being made today on long-term projects — major water resource management activities such as irrigation and hydro-power, drought relief, agricultural land use, structural designs and coastal engineering projects, and energy planning — all based on the assumption that past climatic data, without modification, are a reliable guide to the future. This is no longer a good assumption since the increasing concentrations of greenhouse gases are expected to cause a significant warming of the global climate in the next century. It is a matter of urgency to refine estimates of future climate conditions to improve these conditions.

The implications of such important changes, the conference concluded, "could be profoundly affected by governmental policies on energy conservation, use of fossil fuels, and the emission of some greenhouse gases". The report went on to recommend major studies of the effects on the environment and society of CO₂ and potential ways to control it and other trace gases. After acknowledging that "major uncertainties remain in predictions of changes in global and regional precipitation and temperature patterns", the report concluded: "nevertheless, the understanding of the green-

house question is sufficiently developed that scientists and policy makers should begin an active collaboration to explore the effectiveness of alternative policies".

What, then, is the scientific basis of such bold and pointed conclusions? The next 540 pages of this volume attempt to make the scientific case for the recommendations of the Villach conference. The book presents ten background papers written by well-known scientists for the purpose of summarizing the state-of-the-art in their respective fields. As well as a summary chapter on the greenhouse effect, these include: climatic change and ecosystems issues; emission of CO₂ into the atmosphere; carbon cycle projections of how much CO₂ might remain in the atmosphere; other trace greenhouse gases and aerosols; modelling of the future climate system; empirical studies, analogues and signal detection issues; changing sea level; impacts on ecosystems; relationship to agriculture; and, finally, relationship to global forests.

These chapters are, for the most part, well-presented critical surveys of the areas they consider. Chapter 3, for example, on the carbon cycle, by Bert Bolin, is beautifully written and comprehensive — as we have come to expect from its author. It does not, however, include a discussion of the validity of simple one-dimensional box models for long-term biochemical projections. Although Robert Dickinson's chapter is so comprehensive that I can't easily think of any important climate modelling issue it neglects, it leads off with a differential equation and probably can be fully understood only by analytically trained scientists. But his critical evaluation of the strengths and weaknesses of modelling is first rate, and 'bottom line' valid (even if

downbeat): "The present CO₂ warming signal should lie approximately between 0.3 and 1.1 °C. None of the GCM model results for regional pattern is yet sufficiently certain as to suggest useful regional 'fingerprints' for the warming signal". Similar statements of uncertainty could be quoted from the other chapters, which at the very least provide good up-to-date overviews and excellent reference lists.

As comprehensive as the list of chapters sounds, there are important areas that are not covered. For example, although soil moisture and water supply are mentioned in various chapters, a turn to the index shows that most of the headings containing the word "water" occur in the chapter on sea-level rise, and are not related to the long-term economic planning and user implications of plausible changes to water supply systems. This topic may well prove to be one of the most important issues of trace-greenhouse-gas impacts, and indeed was certainly worthy of a detailed chapter or two to deal, amongst other matters, with the social and political restraints on the rational management of current water supplies. Also missing is coverage of another potentially important issue, the possible direct consequences for health of greenhouse warming: change in disease vectors, morbidity and mortality associated with runs of heat-stress days, and so on. Both of these topics probably deserved their own chapters, although it must be admitted that very little comprehensive work has been done in the latter area, so the editors cannot be faulted for having difficulty in finding people to write about it.

Also in need of more emphasis is a sense of the context of climatic changes for the global economy, and the intrinsic interconnection between increasing levels of trace gases, especially CO₂, and the problem of global economic development. Quite simply, population resources and environmental problems form an integrated set of which climate change is but a part. No nation acting alone can make much difference beyond temporarily slowing down the build-up. Furthermore, any talk of restraint on population size or development in less-developed countries is divisive in the absence of forthright discussion of alternative pathways towards development. This is one of the primary reasons that it is so difficult to achieve a global consensus on how to deal with various trace-gas-producing activities. Developed countries cannot ask the underdeveloped world to re-chart its development expectations simply because of the possibility of effects of global pollution, without at the same time moderating their own profligate consumption of resources. Ultimately, a bargain between rich and poor nations is needed.

One example I have previously discussed — for example, see *The Coevolution*



An accusing finger — the right hand of Grauballe man, whose throat was cut and whose body was buried in the Jutland peat bog c. 1,600 years ago. The unworn fingernails suggest that he was not a manual worker. The picture is taken from European Wetlands in Prehistory edited by John M. Coles and Andrew J. Lawson, just published by Clarendon Press. Price is £50.

of *Climate and Life*, by S.H. Schneider and R. Londer (Sierra Club Books, San Francisco; 1984) — suggests that the rich should control their resource consumption and help put resources on the line to allow the poor to develop, but only if such development is in the context of reducing population growth rates and investing in specific technologies and organizations that have some promise of a sustainable future. To place CO₂ on the global agenda for action outside the context of world development is probably to doom the world to adapt to the changes that inevitably will occur. It would be unfortunate if a potential problem of this magnitude were once again to be handled by random adaptation in the absence of intelligent advanced thinking that could minimize potential damages and even take advantage of potential opportunities.

Despite my dwelling on these omissions, to be fair I should note that the title of the book has "ecosystems" in it, not health or water supply systems. But since the report from Villach was otherwise so comprehensive it is unfortunate that a few more chapters were not added to this already excellent volume.

Finally, the individual authors, especially Dickinson, stress that the problem will alter with time, even though most analyses assume future equilibrium changes. There needs to be a sense of the urgency of the greenhouse problem: both climatic and ecological systems could well be forced to change at virtually unprecedented rates. Therefore, not only will this require study of climatic or biological systems' transient behaviour, but also their coupled transient interactions. It can take centuries, if history is our guide, for forests to march hundreds of kilometres in response to a temperature change of a few degrees. But that magnitude of change could occur in one century or less. How ecosystems can cope with such rates of change, complicated further by other human changes to the physical and chemical environment, is a major uncertainty and is of serious concern. It reminds me of my favourite cliché in dealing with the greenhouse effect issue: "We insult the environment faster than we can understand it or predict the consequences". That point is well documented by Bolin *et al.* But it is a value judgement, of course, whether that is enough to alter global development planning substantially. I myself agree with the Villach participants' report that it is not premature "to explore the effectiveness of alternative policies". Uncertainty is ever present, as SCOPE 29 reiterates; nonetheless, our great biogeophysical experiment continues inexorably into the next millennium. □

Stephen H. Schneider is Head of the Interdisciplinary Climate Systems Section, National Center for Atmospheric Research, PO Box 3,000 Boulder, Colorado 80307, USA.

Open resistance

Julian Davies

Antibiotic Resistance Genes: Ecology, Transfer, and Expression. Edited by Stuart B. Levy and Richard P. Novick. Cold Spring Harbor Laboratory: 1987. Pp.436. \$68.

THE ecological consequences of antibiotic use have provided a prime example of the way in which the environment responds to a global assault with man-made poisons. Human and animal applications of antibiotics have been practised with enormous therapeutic and economic gain since the introduction of the synthetic sulphonamides in the late 1930s; the assault was intensified with the introduction, some ten years later, of penicillin and other secondary metabolic products of microbes. In 1978 alone, 12 million kilograms of antibiotics were produced in the United States — the Earth is literally bathed in a dilute solution of antibiotics.

The response of the endogenous microbial population to this minor catastrophe was unexpected but spectacular: transmissible antibiotic resistance, first identified in Japan in 1959, is now virtually ubiquitous, and is often the principal determinant of how and when antibiotics are used. Watanabe proposed in 1971 that R-plasmids arose through the 'pick-up' of antibiotic resistance determinants by plasmids. Because both plasmids and antibiotic resistance genes were present (independently) in nature well before the antibiotic era, this was a highly plausible notion. With the discovery of transposable elements, and the advent of modern genetic engineering methodology, we have begun to fill in some of the gaps in this molecular history.

In 1985, the Banbury Center of Cold Spring Harbor Laboratory held a symposium, "Antibiotic Resistance Genes: Ecology, Transfer, and Expression", with the specific goal of examining the possible origins of antibiotic resistance genes and assessing the risks associated with their spread in the environment. This is a controversial subject; in spite of good evidence of inter-species dissemination as a result of the wide human and animal use of antibiotics, there are counter-claims that no dramatic changes in antibiotic susceptibility of the Enterobacteriaceae took place in the United States between 1971 and 1984.

The conference embraced three general themes: microbial ecology, gene transfer and maintenance, and evolution of genes and gene transfer systems. The papers, now published in the book under review, are relatively short, and rather uneven in content and in style. Many of the authors

have tended to limit themselves to their own speciality, and although there is a discussion after each paper this too often concentrates on minor points of experimental detail and not on the main issues. The summary discussion session does not always address the questions raised in the introductory chapters, and substantive conclusions are lacking.

In addition to being an object lesson in man's influence on environmental microbiology, antibiotic resistance provides fundamental information for epidemiology, biochemistry and evolutionary science which could aid in guiding the future release of man-made agents (chemicals, microbes) into the environment. The implications for the release of genetically engineered microorganisms (with or without resistance genes) are not discussed in the book. In an area where soundly based biological conclusions might lead to more rationality and less nonsense, it seems that a chance has been missed to present some scientifically valid considerations. For example, plasmid and gene stability are covered, but without reference to broader questions. What determines the movement of genes in what were once considered sterile relationships, such as between Gram-negative and Gram-positive organisms? What is the role of positive selection (of linked characters) versus metabolic load? Such speculations would have added to the value of the book.

Many informative chapters are included, but it is worth noting that several of the authors have also contributed to "Evolution, Ecology and Epidemiology of Antibiotic Resistance", published by Academic Press in 1986. In comparison, the latter has certain advantages and is better focused, but the two together provide an up-to-date and comprehensive overview of what we know about the determinants of antibiotic resistance. Clearly, the last word has not been written on this subject. □

Julian Davies is in the Department of Biotechnology, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris Cédex 15, France.

New in paperback

- *New Guinea Tapeworms and Jewish Grandmothers: Tales of Parasites and People* by Robert S. Desowitz. Publisher is Norton, price is £4.95, \$6.95. For review see *Nature* **296**, 514 (1982).
- *Uneasy Genius: The Life and Work of Pierre Duhem* by Stanley L. Jaki. Publisher is Martinus Nijhoff, price is Dfl. 45, £15. For review see *Nature* **317**, 482 (1985).
- *The History of the Countryside: The Full Fascinating Story of Britain's Landscape* by Oliver Rackham. Publisher is Dent, price is £8.95. For review see *Nature* **321**, 823 (1986).
- *Evolution: Essays in Honour of John Maynard Smith* edited by P. J. Greenwood, P. H. Harvey and M. Slatkin. Publisher is Cambridge University Press, price is £15, \$24.95. For review see *Nature* **319**, 19 (1986).

Whigs, prigs and historians of science

Edward Harrison

The whig interpretation of history, which evaluates the past in terms of the present, is derided by the new historians of science. But their own anti-whig interpretation is priggish and fails to appreciate the temporal depth of scientific research.

PRACTISING scientists have traditionally taken great interest in the history of science and many have contributed scholarly studies (for example, Duhem in mediaeval science, Partington in chemistry, Truesdell in mechanics and Whittaker in physics). In recent decades professional historians of science, trained in the social sciences, have established their own history of science societies, journals and meetings. Natural science and the study of its history have become divergent disciplines, and historical studies now rarely enliven the pages of science journals. Russell, deploring this state of affairs, laments that "Science without its history is like a man without a memory"¹.

Amazon

For our purpose we may say that the history of science resembles a survey of the Amazon river. We chart on the large scale its streams from their sources in the Andes to its complex estuary thousands of miles away, and examine on the small scale the flora and fauna of riverine regions; a comprehensive survey combines both methods. In the history of science, we chart on the large scale streams of ideas from their sources to the complex conceptual scheme of today, and examine on the small scale the individuals and institutions of various periods; a comprehensive survey combines both methods: diachronic and synchronic. Lovejoy's *The Great Chain of Being*, a study of the history of an idea, illustrates the diachronic method; and Westfall's *Never at Rest*, a study of Newton and the affairs of his day, illustrates the synchronic method. History (like the Amazon) evolves; novel sources, insights and fashions of thought re-create the historical vistas, and in a real sense the history of the past lies not behind us but before us.

The new historians of science take to heart Kuhn's maxim², "Insofar as possible . . . the historian should set aside the science that he knows", and learn "from the textbooks and journals of the period he studies". The less the historian knows of the science of today the less he adulterates the science of yesterday. This philosophy, which ratifies ignorance of modern science, runs perilously close to implying that all modern knowledge — not just science — conceals rather than reveals, and that history should be

chronicled by an empty rather than a full mind.

The whig historian, according to Butterfield in his influential *The Whig Interpretation of History*, sees the protestant British constitution as the natural fulfilment of religious and political struggle. The whig interpretation of the past, condemned by Butterfield on several counts, saturates history with value judgements. "The whig historian stands on the summit of the 20th century, and organizes his scheme of history from the point of view of his own day"³. Butterfield cites the example of the man at the street corner who says, "When the Pope ruled England them was called the Dark Ages". The total result of the whig habit of making connections is "to impose a certain form upon the whole historical story, and to produce a scheme of general history which is bound to converge beautifully upon the present". He remarks a few pages later that "the study of the past with one eye, so to speak, upon the present is the source of all sins and sophistries in history . . . it is the essence of what we mean by the word 'unhistorical' ". More specifically, Brush says⁴ that a whig history of science judges "every scientist by the extent of his contributions toward the establishment of modern theories". It leaves unexplored cul-de-sacs (such as the visual-ray, phlogistic and caloric theories) and byways (such as Descartes' principle of action by direct contact and Maxwell's

mechanics of the luminiferous aether), and regards them as impediments in the progress of science.

Rejection

Butterfield's uncompromising rejection of whiggish history stems from his philosophy: the "truth of history . . . is nothing less than the whole of the past, with its complexity of movement, its entanglement of issues, and its intricate interactions, which produced the whole of the complex present"⁵. (Hull remarks⁶ that the truth of history, according to this admirable but impractical definition, consists of a Lewis Carroll map of one mile to the mile.) "Real historical understanding", declares Butterfield⁷, "is not achieved by the subordination of the past to the present, but rather by our making the past our present and attempting to see life with the eyes of another century than our own". According to this proposition, understanding the past precedes rather than follows historical inquiry, which might be true if we had time machines.

Plausibly, the anti-whig interpretation of history derives from the Baconian inductive method, which attempts to investigate phenomena with an observant but empty mind. The deductive whig and inductive anti-whig interpretations are much too simplistic when used exclusively; in combination, however, they provide the essentials of historical inquiry. ►



Joint endeavour — the partnership of Clio the Muse of history (left) and Urania the Muse of astronomy symbolizes the history of science. Each of the Muses jealously guarded her own particular subject and the nine Muses separately and jointly punished all who dared to rival them.

Historians of science who have been taught to dread the whig epithet and are unfamiliar with Butterfield's monograph might suppose that his critique lends support to the maxim: the less we know of modern science the better we can know the science of the past. This would be unjust, for despite his prejudice against whiggery, Butterfield avoids erring grievously in extreme anti-whiggery.

It must be admitted that many scientists, when discussing history, tend to be scandalously careless with quotations and citations, and fail to consult primary sources. They reconstruct situations and make connections with outrageous abandon. They perpetuate whig abridgements of history that attribute too much to too few individuals; for example, more than a millennium of astronomy to Ptolemy and Copernicus, mediaeval technology to Leonardo da Vinci, and pre-seventeenth-century mechanical science to Galileo.

The whig interpretation at one extreme makes a virtue of hindsight and discards from the past what contributes nothing to the present. The anti-whig interpretation at the other extreme makes a virtue of ignorance and discards from the present what contributes nothing to the past. By making a virtue of ignorance, anti-whiggery becomes a form of priggery. I use the word prig to denote narrow-minded superiority, as defined in an essay in an old issue of *The Spectator*⁸ advocating higher education for low-income sections of society: "priggishness is narrow mindedness with a turned up nose".

The whig interpretation of the history of science, practised by most scientists according to historians, commits the crime of reconstructing past science in the context of today's science. The prig interpretation, practised by many historians, adopts a superior attitude to historical work by scientists, and from fear of being unhistorical commits what it supposes is the lesser crime of being unscientific.

Many historians of science are unscientific in the sense that they received their training in the humanities or social sciences; few have gained advanced understanding and carried out research in the modern natural sciences, especially the physical sciences. They concentrate with scrupulous scholarship on a chosen period (the synchronic view) and by narrative discourse depict a landscape with figures. They specialize in the rootless and branchless history of scientists rather than the deeply rooted and many-branched history of science; they psychoanalyse selected individuals (rarely acknowledging that whiggery dictates the selections); they dissect religious and political beliefs; and they document family and social lives, travels, accomplishments, publications and awards.

The prig interpretation defends ignorance of modern science on the grounds

that this knowledge did not exist in the past. Thus a student of mediaeval alchemy need know little modern chemistry. But the reconstructed past is communicated to readers who do not always share the historian's ignorance. Historians who adopt the prig interpretation claim that the exactitudes of modern terminology, such as the difference between heat and temperature, speed and velocity, equilibrium and stability, progress and evolution, did not exist in the period studied. But the historian communicates to an audience in the present and not to the actors in the past,

"The whig interpretation at one extreme makes a virtue of hindsight and discards from the past what contributes nothing to the present. The anti-whig interpretation at the other extreme makes a virtue of ignorance and discards from the present what contributes nothing to the past. By making a virtue of ignorance, anti-whiggery becomes a form of priggery."

and the ears of many members of the audience are attuned to the nuances of modern terminology. The historian may, of course, take refuge from such peer review by publishing in journals normally not read by scientists. By communicating the history of science only to historians, however, and deigning not to communicate it to scientists, the historian lives neither in the past nor the present, in a never-never land where ignorance is bliss, unable to evaluate (and if necessary to discount) the effect of modern science on modern styles of thought.

Reconstructing the past requires vigilant commentary on differences between past and present sciences and languages. Abandoning the vantage point of modern science while retaining the advantage of modern language distorts the historical reconstruction. How will the future historian, studying the twentieth-century history of pre-twentieth-century science, allow for the ignorance of present-day historians of science?

Professional historians of science, possibly as a result of their training in history, tend to view the past in terms of biographies of individuals and activities of institutions; scientists tend to view the past as a historiographic network of inter-related experiments and theories. The historian recalls the synchronic events of a particular period; the scientist stands back, looks at the same period, and sees ideas diachronically cascading from earlier periods, forming a tumult and having an effect on later periods. There seems no reason why a history of science cannot combine both views.

Without some compromise between historians and scientists — between anti-

whiggery and whiggery — recent history of science must suffer from lack of professional attention. An anti-whig historian who knows little modern science will skip this period and explore earlier and much safer periods; a scientist who fears the opprobrium associated with the whig epithet will also avoid this period and write only reviews of current developments. Fortunately for the history of science, most scientists remain unconcerned with the philosophical intricacies of historical inquiry, and readily discuss the science of earlier decades. Their occasionally inflated recollections of personal contributions, usually discounted by colleagues, cause comparatively little harm. Better to have eye-witness accounts than historians' guesses centuries later.

Traditionally, scientists have considerable interest in the history of science. When physicists seek to justify the modern quest for unified interactions, they naturally turn to the age of Maxwell where the quest began with the unification of electricity and magnetism. The notion that scientists have little interest in the origins of their subject is of recent vintage, concocted, one suspects, by historians defending new territory.

In scientific research, where ideas form and dissolve in a state of flux and at any moment present countless potential futures, scientists retain their bearings by contrasting past and present ideas. Awareness of temporal depth in science forms an integral part of scientific research. The references in any scientific work constitute the visible evidence of its relations with preceding scientific work. Scientists grope their way, seeking to divine where they are going from where they are coming; they reach into the future as well as the past, and their diachronically extrapolated conceptual networks are not whiggish sins but the essence of science in action.

Historians are right: scientists need instruction on how to examine the past in a scholarly and objective manner; but historians are wrong: scientists cannot be excluded from the history of science, for temporal depth is an integral part of scientific research. □

1. Russell, C. "Whigs and professionals" *Nature* **308**, 777–778 (1984).
2. Kuhn, T. "The history of science" in *International Encyclopedia of the Social Sciences* Vol. 14, 74–83 (Macmillan, New York, 1968).
3. Butterfield, H. in *The Whig Interpretation of History* 13, 12, 31–32 (Bell & Sons, London, 1968).
4. Brush, S.G. "Should the history of science be rated X?" *Science* **183**, 1164–1172 (1974).
5. Butterfield, H. in *The Whig Interpretation of History* 19 (Bell & Sons, London, 1968).
6. Hull, D.L. "In defense of presentism," *Hist. Theor.* **18**, 1–15 (1979).
7. Butterfield, H. in *The Whig Interpretation of History* 16 (Bell & Sons, London, 1968).
8. Anon. "Social settlements" *The Spectator* 267–268, 19 Feb. 1898.

Edward Harrison is a Professor in the Department of Physics and Astronomy, University of Massachusetts, Amherst, Massachusetts 01002, USA.

New optimization methods from physics and biology

David G. Bounds

Royal Signals and Radar Establishment, St Andrews Road, Malvern, Worcestershire WR14 3PS, UK

Research in spin-glass physics, population genetics, and neural network dynamics has provided powerful methods for finding near-global optima of functions that have many local optima. These techniques are being applied successfully to a wide variety of scientific and engineering problems. They may also give new insights into combinatorial optimization problems.

MANY problems in science and engineering can be formulated as optimization problems in which some cost function must be minimized. Many interesting optimization problems are hard because they are *NP* complete, that is an exact solution requires a number of computational steps that grows faster than any finite power of some appropriate measure of the size of the problem. A classic example is the 'travelling salesman problem' which consists of finding the shortest tour between N cities visiting each once only and ending at the starting point. For N cities there are $(N-1)!/2$ possible tours. As the number of tours grows faster than any finite power of N , it is widely believed (though not proved) that travelling salesman problems require an exponential time to compute¹.

But, sales managers do not require that their salesmen choose the perfect route; only a reasonable one. Usually near-optimal solutions are satisfactory, if such solutions exist. There are grounds for believing that at least some *NP*-complete problems do have many local optima with values close to the global optimum (see Fig. 1). In these problems there is little to choose between the many solutions and any one of them will be a

reasonable estimate of the global optimum value. Furthermore, in many engineering applications the choice of cost function is not unique. In such cases it is more important to find good solutions quickly than to find the global minimum.

Three pieces of work in apparently unrelated areas of physics and biology seem likely to lead to much better algorithms for finding acceptably low-cost minima; and certainly suggest novel ways for thinking about such problems. The first, optimization by simulated annealing, originated in spin-glass physics. The second was stimulated by neurophysiology and models of neural networks. The third is based on ideas from biological evolution. All three methods have been applied to the travelling salesman problem which has the advantage that it is a well-characterized problem that provides a good testing ground. It is important, however, to note that the new techniques described here are not limited to combinatorial problems; they are also applicable to continuous functions. Indeed, in the neural network optimization described later much of the power arises from mapping a discrete space travelling salesman problem into a continuous space. Thus although this review deals mostly with the discrete travelling salesman problem to provide a concrete example, most methods are widely applicable to continuous domains. Packel and Traub have recently reviewed the issue of computational complexity for continuous problems which contain only part of the information necessary for a complete solution². This kind of computational, information-based complexity, is characteristic of many real-world applications but it will not be considered here.

Spin-glass problem

The theory of spin-glasses is an active field of research from which some unusual new physics has emerged^{4,5} principally from studies of the Sherrington-Kirkpatrick model⁶. A dilute, disordered magnetic material may exhibit a low-temperature phase in which the thermal energy is not sufficient to prevent spins freezing into particular orientations, but the competing magnetic interactions are too weak to enforce the long-range order of a ferromagnet. This frozen disordered state is believed to be the structure of the spin-glass phase. An essential feature of spin-glass models is the idea of frustration³ which arises because it is not possible to find a spin configuration which satisfies all bond constraints between spins.

The current picture would probably not have emerged without extensive numerical simulations which have guided the development of analytical theories. The principal tool for numerical work is the Monte Carlo method due to Metropolis *et al.*⁷. A central question in these systems is whether or not some freezing transition to a spin-glass phase really occurs as the temperature is lowered. To study this, one carries out Monte Carlo

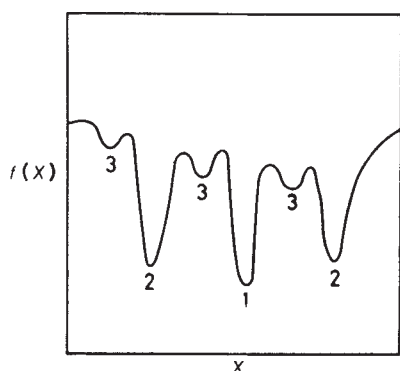


Fig. 1 A two-dimensional representation of a cost function $f(X) = f(x_1, x_2, \dots, x_N)$ which has many local minima with values of f close to the global minimum value. X defines an allowed configuration. In the N -city travelling salesman problem $f(X)$ is the length of a tour and X is a N -dimensional vector which defines in order that cities occur in a tour; for example the permutation $P = (i_1, i_2, \dots, i_N)$ where i labels the cities. Point 1 is the global minimum (shortest tour); points 2 are low-cost local minima which are good enough for many practical problems (short tours); points 3 are high-cost local minima from which minimization methods must be able to escape. Note that the form of $f(X)$ depends on the choice of move set since the move set defines which points are close in X .

calculations at a range of temperatures. A time-varying temperature $T(t)$ is an 'annealing schedule', and the method seeks the minimum of the cost function (hamiltonian) with respect to flips of the spins $\sigma = \pm 1$:

$$H = -\sum_{i=1}^N \sum_{j=i+1}^N J_{ij} \sigma_i \sigma_j \quad (1)$$

Kirkpatrick *et al.*⁸ have proposed the same method as a general optimization technique to find the minimum values of arbitrary cost functions, and the method is now known as optimization by simulated annealing (OSA). It has been used successfully for a number of hard optimization problems arising in computer chip design, computer vision, and the travelling salesman problem (for which it was proposed independently by Černý⁹). The connection between the Metropolis method and minimization was noted earlier by Pincus¹⁰.

The OSA method operates thus. An optimization problem involves a cost function $f(X) = f(x_1, x_2, \dots, x_N)$ to be minimized and a set of candidate solutions generated by trial moves. Gradient descent methods accept only those moves which reduce $f(X)$ and therefore cannot escape from local minima. OSA accepts trial moves that decrease $f(X)$ but also allows some moves which increase $f(X)$. If a trial move increases $f(X)$ by $\Delta f(X)$, it is accepted with probability $\exp[-\Delta f(X)/T]$ where T is a control parameter (Boltzmann's constant multiplied by the temperature for a spin-glass—hence the term 'temperature' in the OSA literature regardless of what cost function is involved).

A system evolving under these rules eventually reaches thermal equilibrium at any given temperature, and the relative probabilities of two states α and β is given thereafter by the Boltzmann distribution

$$(P_\alpha/P_\beta) = \exp[-\{f(X_\alpha) - f(X_\beta)\}/T] \quad (2)$$

where P_α is the probability of being in state α of cost $f(X_\alpha)$. At high temperatures the probability of accepting moves that increase $f(X)$ is greater and equilibrium is reached more quickly. At low temperatures equilibration takes longer but the system is more heavily weighted towards low-cost states. The strategy therefore is to do a coarse search of the space at high temperature and then reduce T to focus on the low-cost states. Apart from designing a suitable cost function, the only difficulty in applying OSA is to choose an efficient annealing schedule which nevertheless ensures that the system has a Boltzmann distribution of states down to as low a temperature as possible (within a given computer budget). To do this, it is necessary to keep the system close to equilibrium throughout the annealing process. The rate of cooling can be quite critical, as the physical analogy on which OSA is based suggests. As metallurgists know well, it is necessary to cool a melt slowly near the freezing temperature to grow near-perfect crystals (which correspond to global minimum states). Very rapid cooling, by splat quenching, results in amorphous structures (metastable states); while lowering T too slowly may result in supercooling if the melt fails to nucleate (that is, the system approaches the ground state very slowly). While some rigorous bounds on the rate of annealing have been proposed^{11,12}, most work has used more physically motivated criteria such as monitoring the specific heat¹³, $C(T)$:

$$C(T) = [\langle f(X)^2 \rangle_T - \langle f(X) \rangle_T^2] / T^2 \quad (3)$$

An increase in $C(T)$ signals the onset of a phase transition whence a slower cooling rate is necessary.

Several studies of the travelling salesman problem have been carried out using OSA^{8,9,14,15}. Most studies make use of Lin's move set¹⁶ where a tour is specified by the permutation $P = (i_1, i_2, \dots, i_N)$ with associated length

$$L = d_{i_1 i_2} + d_{i_2 i_3} + \dots + d_{i_N i_1}$$

d_{ij} being the distance between cities i and j . A tour is said to be ' λ -OPT' if no shorter tours can be obtained by replacing λ steps of the tour (bonds) with any other set of λ bonds. This

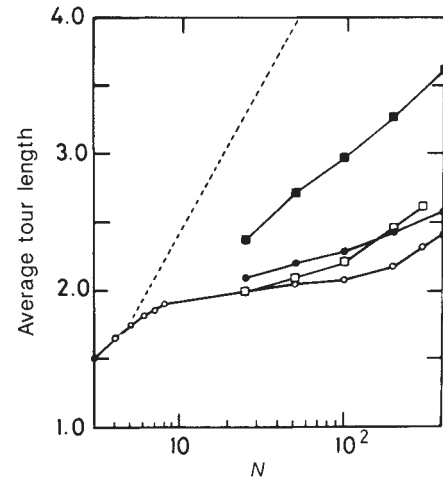


Fig. 2 TSP results from Kirkpatrick and Toulouse¹⁴. This figure shows the optimal tour length for random distance travelling salesman problems of up to $N = 400$ cities, as obtained from various algorithms. The dashed line is the upper bound provided by the greedy algorithm. ("Choose a city at random. The next city on the tour is the closest one".) The solid points are from iterative improvement using exhaustive search for two-bond rearrangements (squares) and three-bond rearrangements (dots). The open points for $N > 24$ are OSA results using two-bond moves (squares) and three-bond moves (circles). The open data for $N \leq 12$ are exact results. Reproduced with permission of the publisher.

provides a natural set of increasingly restrictive constraints to satisfy as: (1) a λ -OPT tour is also λ' -OPT for $\lambda' < \lambda$ ¹⁵; (2) an N -city tour is optimal if and only if it is N -OPT.

One simple move is the interchange of two bonds: two steps $d_{i,i+1}$ and $d_{j,j+1}$ on the tour are replaced by $d_{i,j}$ and $d_{i+1,j+1}$. It has been found that for $N > 50$, the OSA technique compares favourably with exhaustive two-bond and three-bond searches, and is progressively better than gradient descent at finding short tours as N increases (Fig. 2).

In addition to providing a useful simulation tool, spin-glass physics may yield deeper insights into combinatorial optimization problems and computational complexity. The justification for applying spin-glass methods to NP -complete problems, such as the travelling salesman problem, is that the central problem of finding the ground-state energy of the three-dimensional spin-glass is NP complete^{17,18}. Several workers have formulated the travelling salesman problem in statistical mechanical terms. Orland¹⁹ has conjectured that NP completeness is associated with replica symmetry breaking (a much used technique in spin-glass physics where it was introduced by Edwards and Anderson²⁰; see ref. 5 for a review of the replica method). A connection between replicas and P -class optimization problems has also been suggested by Mezard and Parisi²¹, and their results have recently been extended to the NP case (M. Mezard and G. Parisi, in preparation). It remains to be seen whether these rather technical connections are valid. When the travelling salesman problem is transformed to a statistical physics problem, quantities such as the average length of tours become thermodynamic quantities, and one is interested in their behaviour as a function of temperature. Vannimenus and Mezard²² have shown the existence of two different temperature regimes which exhibit quite different behaviour in the way such properties vary as a function of N . They have also obtained the exact asymptotic behaviour of the optimal tour length; though only for a travelling salesman problem in infinite dimension. Kirkpatrick and Toulouse¹⁴ have introduced a simplified version of the travelling salesman problem that may prove analytically soluble, as did the Sherrington-Kirkpatrick model for spin-glasses. This is a fascinating area and the reader is urged to consult the pioneering

paper¹⁴ for a deeper discussion. Among the tantalizing possibilities is that a more refined characterization of computational complexity may emerge.

Returning to optimization techniques, the great advantages of OSA are simplicity and generality. Though it may need ingenuity to choose an appropriate cost function and move set, the method is simple to implement and has been used successfully on a variety of problems. The main disadvantage is that choosing an annealing schedule for practical purposes is still something of a black art.

Networks of artificial neurons

OSA is essentially a serial computation. To satisfy the requirements of detailed balance, which is necessary if a Boltzmann distribution is to be attained, one trial move must be accepted or rejected before another can be tried. Only in a few special cases—which are of great interest to physicists but not typical of other optimization problems—can multiple moves be made simultaneously thus enabling the method to be implemented efficiently on powerful parallel computers. It has often been suggested that parts of biological perception tasks such as the early stages of vision and speech perception can be formulated as optimization problems²³. In view of the massive amounts of data to be processed at any instant, and the rapidity with which solutions to such highly complicated problems are found, it is clear that biological information processing systems must carry out highly parallel computations. Can simplified models of neural networks be made which retain this computational power and speed for other optimization problems?

Although the mechanisms of biological computation remain largely a mystery, recent work by Hopfield and Tank²⁴ suggests that the high connectivity of neural networks (each neuron is connected, typically, to 10^2 – 10^4 others) together with the fact that neurons are more akin to nonlinear analog devices than to binary switches, are key ingredients. Hopfield and Tank have shown²⁴ that a network of nonlinear analog devices (artificial neurons) working in parallel can give good solutions to *NP*-hard optimization problems within a few timesteps, where time is measured in the characteristic response time of the neurons. For real neurons the characteristic times are in the millisecond range while for electronic 'neurons' the time constants could be much faster—perhaps a few nanoseconds. The fact that analog computation is less accurate than digital computation is unimportant for hard optimization problems because one is seeking good approximations to the global minimum rather than the exact value.

The basic element (neuron) is an amplifier with a nonlinear response function which relates the output voltage, V , to the input voltage, u . The form of the response function is shown in Fig. 3. Details of the circuitry would take too long to describe here; they are given in ref. 24. The relationship of these electronic circuits to biological neural networks is described at length in ref. 25. Initially, circuits were simulated by integrating the appropriate equations of motion²⁴. But future applications will depend upon building circuits in hardware rather than solving the corresponding differential equations on a serial computer. Some prototype chips have been produced already^{26,27}.

It is by no means obvious how optimization problems can be mapped onto networks like these. For the travelling salesman problem, Hopfield and Tank²⁴ represented tours by permutation matrices, which requires a network of N^2 neurons for N cities. Hopfield and Tank tested the method on a 10-city travelling salesman problem. The evolution of the network to a stable solution is shown in Fig. 4. For a chosen set of initial conditions the calculation is deterministic, but different initial conditions lead to different solutions. In 20 simulations starting from different initial states, Hopfield and Tank found that 16 converged to valid tours, and about half of the trials produced one of the two shortest tours found by exhaustive search. This is an impressive performance: for 10 cities there are $(N-1)!/2 =$

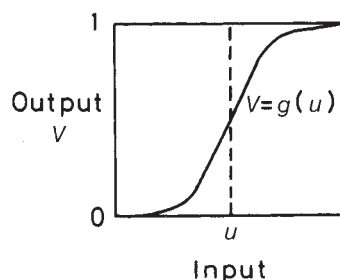


Fig. 3 Response function $V = g(u)$ for artificial neurons (amplifiers).

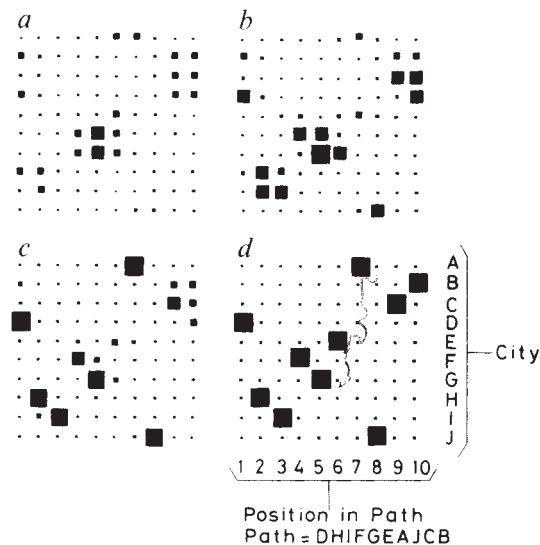


Fig. 4 Hopfield and Tank's travelling salesman results²⁴. This shows the convergence of the 10-city analog circuit to a tour. The linear dimension of each square is proportional to the probability of city i occurring at position j in the tour. Panels *a*, *b* and *c* show intermediate times in sequence; *d* is the final state, which corresponds to cities A...J being visited in the order DHIFGEAJCB.

Reproduced by permission from ref. 24.

181,440 different tours. However, it will require simulations of much larger networks to know if the method will prove useful for real optimization problems or if, instead, good solutions are strongly dependent on the initial conditions. Preliminary calculations on a 30-city travelling salesman problem²⁴ suggest that this may be a problem. There is also much work to be done to discover the range of computations for which such networks are efficient. A number of applications are described in ref. 25.

Recently another parallel algorithm for the travelling salesman problem which also has its roots in modelling neural network behaviour (in this case topographically ordered projections in the brain), has been reported by Durbin and Willshaw²⁸. In this case a 'rubber band' is gradually stretched until it passes arbitrarily close to all cities on the tour. This method is claimed to produce shorter tours than the Hopfield and Tank algorithm, to scale well with problem size; and to be extendable to many optimization problems²⁸.

Genetic algorithms

Evolution provides perhaps the ultimate example of a complex optimization problem; though not one for which a plausible cost function can be written down. The idea of studying adaptation mechanisms in biosystems rigorously and of using similar

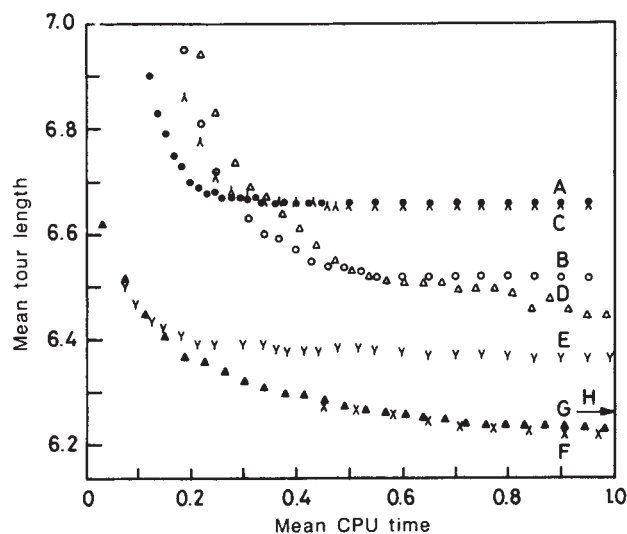


Fig. 5 Brady's results for the travelling salesman problem³⁰. A comparison of eight different optimization strategies. In each case, data were averaged over 100 separate optimizations. A, quenching by random attempted mutations. B, the best of two independent quenches. C, competition between two quenches. D, mating between two quenches. E, mating between two quenches obtained by systematic search. F, mating between 12 quenches obtained by systematic search. G, the best of n independent quenches obtained by systematic search, $n = 1, 2, 3, \dots, 26$. H, simulated annealing; 20,000 random attempted mutations per temperature (statistics from 20 optimizations only, with mean time 5.6 s). Reproduced by permission from ref. 30.

strategies for complex optimization problems was pioneered by Holland²⁹. These methods are now called genetic algorithms. Although there has been quite a lot of activity in this field, little has appeared in the literature. The recent application of similar techniques to the travelling salesman problem by Brady³⁰ provides an illustration.

Consider an arbitrary tour of N cities. In biological terminology a random alteration to the tour is a mutation. Brady restricted his study to local mutations obtained by swapping the order of two cities in the tour. Such mutations belong to a subset of Lin's two-bond moves which were used in the annealing studies described earlier. The baseline on which all other methods improve (method A, Fig. 5) is to choose trial mutations but accept only those which shorten the tour. This single random quench is merely an inefficient way of finding the nearest local minimum, and Brady suggested that a lineage developing in a similar manner would be at an evolutionary disadvantage because it is destined to stagnate. Prospects of survival are improved by diversification into several independent species, but evolution may progress more slowly as resources must be shared. For optimization, this suggests dividing the available computer time into several independent starting tours and quenching each separately (method B). The effect of competition between species may be modelled by taking two independent tours, carrying out random quenches, but for every τ attempted mutations the longer tour is discarded, a second copy of the shorter tour is made, and the quenches are continued (method C). As this forces the two lines back to a common point, it is not surprising that on average this method performs no better than single random quench. A more promising variant is suggested by sexual species where genes are swapped during mating. This translates to a strategy (method D) similar to method C but instead of discarding all of the longer tour after τ attempted mutations, the tours are searched for subroutes where some common subset of cities are visited in a different order (see Fig. 6). The shorter subroute is then copied over to replace the

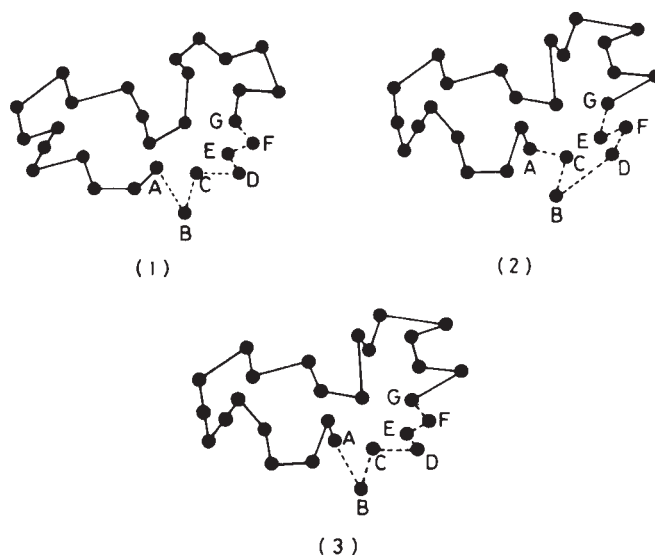


Fig. 6 Gene swapping during mating. Panels (1) and (2) show two independent tours after τ mutations. The subset of cities (A...G) occurs in both tours, but they are visited in a different order. Because the subroute ABCDEFG is shorter than ACBDFEG, the dashed segment of tour (2) is replaced with that of tour (1) to give a new tour (3). Tour (2) is discarded and optimization continues from (1) and (3).

longer one. One would expect this method to produce shorter tours than A and C; and also to be some improvement on B, which has two independent paths, as some of the best features of both paths are included in D. This is observed.

The last two methods used by Brady take the mating strategy of method D but replace the random quench with a systematic search for favourable mutations. In method E, mating is only carried out between two independent tours when each is optimum in the sense that all pair interchanges of cities would give longer tours. Mating then provides an escape route from local minima. Method F is very similar to method E but there are twelve independent paths instead of two. Method F produces the shortest tours found, and Brady suggests that it is more successful than mating between a single pair of individuals (method E) because there is a wider gene pool from which to choose subroutes.

On the limited numerical data available it is not certain that method F performs better than OSA for 64-city travelling salesman problems. Brady carried out 100 different runs for each genetic method, and 20 short OSA runs for a single configuration of 64 cities. To reach any firm conclusions would require averaging not only over a larger number of runs for a given instance of the 64-city travelling salesman problem, but over many different instances as well. In spin-glass terms, it is vital to average over bond sets $\{J_{ij}\}$ to obtain quantitative information. Furthermore, even if one of these methods does perform better than OSA for 64 cities, it is not certain to do so for larger problems.

It is worth observing that different mating strategies correspond to different move sets and therefore alter the cost function surface (see Fig. 2). It is possible in principle that a move set might be found that turns a function with multiple minima into a near-convex one. A thorough comparison of simulated annealing, genetic algorithms, and neural network methods would be very interesting.

Applications and prospects

The optimization methods outlined here are still in their infancy, but they have much to offer. OSA has already been applied to a wide range of practical problems from the layout of com-

ponents on very-large scale integrated circuit chips to rerouting garbage collection trucks in Grenoble. The network approach proposed more recently by Hopfield and Tank²⁴ may prove equally valuable for repetitive tasks where it is worthwhile to build special purpose hardware, and genetic algorithms will receive more attention as parallel SIMD (single instruction, multiple data) and MIMD (multiple instruction, multiple data) computers become more widely available.

The recent history of optimization techniques provides a vivid

example of how ideas, developed in apparently unrelated fields, can have a great impact in new areas. Not long ago, few would have predicted that the theoretical study of spin-glasses, evolution, and neural network modelling would provide economical routes for travelling salesmen.

I thank the authors of refs 14, 24 and 30 for permission to reproduce figures from their papers, and J. J. Hopfield, D. W. Tank, R. M. Brady, M. Mezard, D. Willshaw and R. Durbin for copies of their work before publication.

1. Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G. & Shmoys, D. B. *The Travelling Salesman Problem* (Wiley, New York, 1985).
2. Packer, E. W. & Traub, J. F. *Nature* **328**, 29–33 (1987).
3. Toulouse, G. *Comm. Phys.* **2**, 115–119 (1977).
4. Van Hemmen, J. L. & Morgenstern, I. (eds) *Lecture Notes in Physics* Vol. 192 (Springer, 1983).
5. Binder, K. & Young, A. P. *Rev. mod. Phys.* **58**, 801–976 (1986).
6. Sherrington, D. & Kirkpatrick, S. *Phys. Rev. Lett.* **32**, 1972 (1975); *Phys. Rev.* **B17**, 4384–4403 (1978).
7. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. *J. chem. Phys.* **21**, 1087–1092 (1953).
8. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. *Science* **220**, 671–680 (1983).
9. C rny, V. *J. Optimisation Theory Appl.* **45**, 41–51 (1985).
10. Pincus, M. *Optimisation Res.* **18**, 1225–1228 (1970).
11. Geman, S. & Geman, D. *IEEE PAMI* **6**, 721–741 (1984).
12. Gidas, B. *J. stat. Phys.* **39**, 73–131 (1985).
13. Kirkpatrick, S. *J. stat. Phys.* **34**, 975–986 (1984).
14. Kirkpatrick, S. & Toulouse, G. *J. Phys., Paris* **46**, 1277–1292 (1985).
15. Bonomi, E. & Lutton, J. *SIAM Rev.* **26**, 551–567 (1984).
16. Lin, S. *Bell Syst. Tech. J.* **44**, 2245–2269 (1965).
17. Bachas, C. P. *J. Phys.* **A17**, L709–L712 (1984).
18. Barahona, F., Maynard, R., Rammal, R. & Uhry, J. P. *J. Phys.* **A15**, 673 (1982).
19. Orland, H. *J. Phys. Lett., Paris* **46**, L763–L770 (1985).
20. Edwards, S. F. & Anderson, P. W. *J. Phys.* **F 5**, 965–974 (1975).
21. Mezard, M. & Parisi, G. *J. Phys. Lett., Paris* **46**, L771–L778 (1985).
22. Vannimenus, J. & Mezard, M. *J. Phys. Lett., Paris* **45**, L1145–L1153 (1984).
23. Poggio, T., Torre, V. & Koch, C. *Nature* **317**, 314–319 (1985).
24. Hopfield, J. J. & Tank, D. W. *Biol. Cyb.* **52**, 141–152 (1985).
25. Hopfield, J. J. & Tank, D. W. *Science* **233**, 625–633 (1986).
26. Graf, H. P. et al. *AIP Conf. Proc. 151 Neural Networks for Computing*, 182–187 (1986).
27. Sivilotti, M., Emerling, M. R. & Mead, C. A. *AIP Conf. Proc. 151 Neural Networks for Computing*, 408–418 (1986).
28. Durbin, R. & Willshaw, D. *Nature* **336**, 689–691 (1987).
29. Holland, J. H. *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, 1975).
30. Brady, R. M. *Nature* **317**, 804–806 (1985).

Functional inactivation of genes by dominant negative mutations

Ira Herskowitz

Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California 94143, USA

Molecular biologists are increasingly faced with the problem of assigning a function to genes that have been cloned. A new approach to this problem involves the manipulation of the cloned gene to create what are known as 'dominant negative' mutations. These encode mutant polypeptides that when overexpressed disrupt the activity of the wild-type gene. There are many precedents for this kind of behaviour in the literature—some oncogenes might be examples of naturally occurring dominant negative mutations.

AN awe-inspiring array of techniques is now available for the cloning of genes. For example, the cloned gene may be isolated on the basis of its pattern of expression¹, or by its ability to hybridize to heterologous probes or to code for a particular antigen². The nucleotide sequence may be revealing if it leads to identification of the protein product of the gene and to a suggestion of the general class to which the protein belongs—for example, a DNA-binding regulatory protein, a protease, or a protein kinase. In general, however, the problem of identifying the function of the gene—what its protein actually does in a living cell—will remain. The classical approach is to inactivate the gene and see what effects this has. The function of a gene can, in principle, be inactivated at any of several levels. In yeasts and other lower eukaryotes the wild-type gene can itself be disrupted by targeted insertion, but as yet it is not possible to achieve this in higher eukaryotes, which has been a major stumbling block in determining the functions of mammalian genes that contain the 'homeobox' sequences characteristic of the homoeotic genes of *Drosophila*. One approach that has been suggested involves the disruption of gene function at the level of RNA, but as yet this has met with mixed success.

The purpose of this article is to describe a new strategy in which the function of a gene is blocked at the protein level: the

cloned gene is altered so that it encodes a mutant product capable of inhibiting the wild-type gene product in a cell, thus causing the cell to be deficient in the function of that gene product. Such a mutation would be 'dominant' because its phenotype is manifested in the presence of the wild-type gene, and as it inactivates the wild-type gene function it is referred to as a 'dominant negative' mutation. Muller³ referred to mutations of this kind as 'antimorphs'—"antagonistic mutant genes, having an effect actually contrary to that of the gene from which they were derived". A dominant negative mutation is to be contrasted with other types of dominant mutation, such as those that increase the activity of the gene product—'hypermorphs' in Muller's terminology. Before describing the new strategy I shall review other methods currently in use to inactivate genes in eukaryotes, as a basis for a comparison of the different strategies.

Current strategies

Gene disruption. In some organisms a wild-type gene can be replaced by a mutant version by homologous recombination. The principal virtue of this method is that the gene function is completely destroyed, but at present the technique is only applicable to yeasts^{4–6}, *Aspergillus*^{7,8} and *Dictyostelium*⁴⁴. In higher

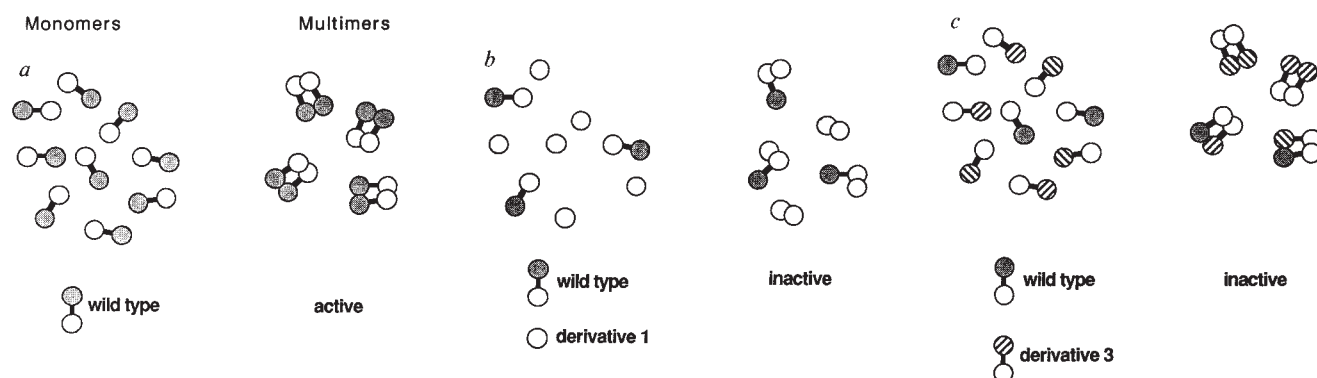


Fig. 1 Inhibitory polypeptides that interfere with the function of multimeric proteins. The polypeptide chain shown here contains two domains, amino-terminal (open) and carboxy-terminal (shaded). *a*, Wild-type polypeptides associate through their amino-terminal domains (as does, for example, the *trp* repressor; see text) to form an active dimeric protein. *b*, Derivative 1 codes for the amino-terminal domain and is able to associate with wild-type chains. When derivative 1 polypeptides are produced in excess, only inactive multimers are formed. *c*, Derivative 3 contains multiple alterations in its carboxy-terminal domain. Association of such monomers with wild-type subunits also leads to formation of dimers that are functionally inactive. Formation of derivatives 1 and 3 is described in Fig. 3 and in the text.

eukaryotes, DNA segments can be added to the genome in random positions, but it is not possible to inactivate the equivalent wild-type gene by targeted homologous recombination. Some examples of homologous recombination involving exogenous DNA introduced into mammalian tissue culture cells have been reported^{9,10}, but their occurrence is rare.

Antisense RNA. An alternative approach that has attracted much attention involves the use of 'antisense' RNA to block the expression of a gene by preventing the translation of its sense transcripts^{11,12}. There have been some notable successes with this method of producing 'mutants without mutations'¹³: for example, a phenocopy of *Drosophila Krüppel* mutants has been produced by injecting antisense *Krüppel* RNA into wild-type embryos¹⁴, a discoidin-deficient slime-mould results from expression of antisense discoidin RNA *in vivo*¹⁵, and antisense RNA of the proto-oncogene *c-fos* can inhibit the platelet-derived growth factor (PDGF)-induced growth of mouse 3T3 cells that express it¹⁶. Although further successes may result from refinements of this method^{12,17-19}, the antisense approach has not yet led to the flood of information that I, for one, expected would result from its use.

Imbalance. Meeks-Wagner and Hartwell²⁰ have reported the creation of a mutant phenotype by overexpression of a wild-type gene. The rationale is that an imbalance of subunit concentration can have dire consequences for the proper formation of multi-protein structures, such as the cytoskeleton and the histone scaffolding of eukaryotic chromosomes. In the particular example described, an altered ratio of histones H2A-H2B to H3-H4 resulting from overexpressing the genes for histones H2A-H2B was apparently the cause of chromosome loss.

Antibodies. Gene function can also be inhibited by the use of antibodies against synthetic antigens based on the predicted sequence of the gene product. For example, antibodies against microtubule components can alter the morphology of intermediate filaments^{21,22}, and antibodies against *ras* proteins can transiently reverse the transformed phenotype of *ras*-transformed cells^{23,24}. Although suitable for the analysis of individual cells or even groups of cells, this method is limited to transient perturbation of the wild-type gene function because of antibody dilution and degradation.

Dominant negative mutations

This new approach involves the inhibition of the function of a wild-type gene product by an overproduced inhibitory variant of the same product. Such inhibitory variants of a wild-type product can be designed because proteins have multiple functional sites that can be mutated independently: for example, sites for oligomerization, substrate-binding, catalysis, membrane

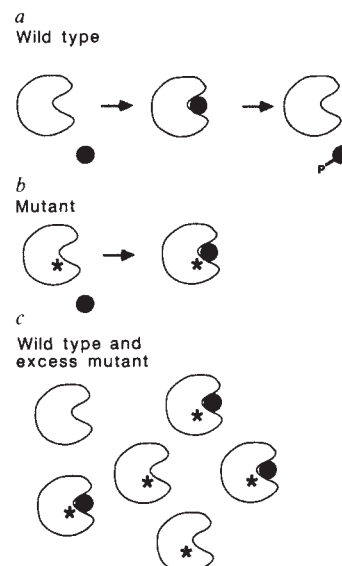


Fig. 2 Inhibitory polypeptides that interfere with function of monomeric proteins. *a*, The enzyme shown is a monomeric enzyme, for example, a protein kinase, which phosphorylates its substrate (black circle). *b*, The mutant form of the enzyme contains an amino-acid substitution that inactivates the catalytic site of the enzyme (asterisk) but that does not affect binding of substrate. *c*, The mutant enzyme acts as a competitive inhibitor of the wild-type enzyme under conditions when the mutant enzyme is overproduced and when substrate is limiting. For example, overproduction of a defective form of the tyrosine kinase coded by proto-oncogene *c-src* might result in a cell depleted for this kinase activity.

association, and so on. Thus, if a protein is multimeric, a derivative capable of interacting with wild-type polypeptide chains but otherwise defective will be inhibitory if it causes the formation of non-functional multimers (Fig. 1). It may also be possible to form inhibitory derivatives of monomeric proteins: if, for example, the activity of a protein is limited by the availability of substrate, then a variant capable of binding substrate but not of carrying out a subsequent catalytic step could be inhibitory (Fig. 2).

In general, therefore, dominant negative mutant proteins will retain an intact, functional subset of the domains of the parent, wild-type protein, but have the complement of this subset either missing or altered so as to be non-functional. It cannot, of course, be known in advance precisely how such a mutant can be created starting with an open reading frame obtained from

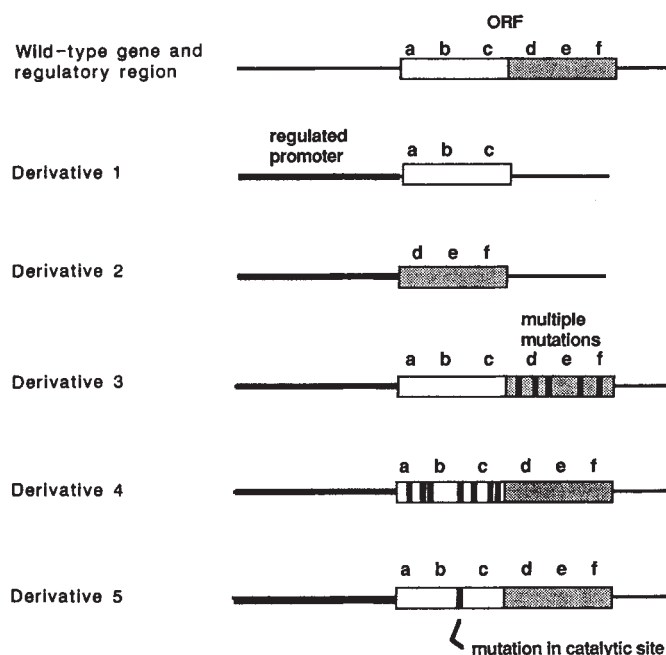


Fig. 3 Structure of a wild-type gene and derivatives designed to overproduce an inhibitory polypeptide. The top line shows the structure of a cloned segment containing an open reading frame (ORF) designated a-f which is read rightwards. The five derivatives all contain a regulatory region upstream of the ORF that confers regulated overproduction of the modified polypeptide coded by the ORF. Transcription is activated, for example, by the addition of steroid hormone or metal ion for the corresponding regulatory regions, or can be governed by tissue-specific regulatory regions. The five derivatives contain alterations in the ORF with the hope that at least one of these modified polypeptides will be capable of interfering with the function of wild-type subunits already in the cell (see Figs 1 and 2). Derivative 1 codes for an amino-terminal segment of the wild-type polypeptide; derivative 2 codes for a carboxy-terminal segment of the wild-type polypeptide; derivatives 3 and 4 code for polypeptides with several amino-acid substitutions in the carboxy- or amino-terminal segments as indicated. If something is known about the function of the polypeptide, an 'educated' derivative (derivative 5) can be designed. For example, the one shown here codes for a polypeptide chain altered for an amino-acid residue that is essential for catalytic activity (as in Fig. 2).

the cloned gene under investigation. However, I suggest that the following four derivatives would give a good chance of creating at least one dominant negative polypeptide (see Fig. 3): derivative 1, which encodes an amino-terminal segment of the polypeptide; derivative 2, which encodes a carboxy-terminal segment of the polypeptide; derivative 3, which is a full-length polypeptide, but with an extensively mutagenized carboxy-terminal segment, and derivative 4, which is analogous to derivative 3, but with the amino-terminus mutagenized. If the enzymatic function of the gene product is known or can be inferred then it might be possible to make specific mutations to generate an additional derivative (derivative 5).

The expression of these derivatives would be put under the control of a regulated promoter, such as the glucocorticoid response element (GRE) of mouse mammary tumour virus, which would confer inducibility by glucocorticoids²⁵ or a tissue-specific regulatory element²⁶. This would have the important consequence that the mutant phenotype would be conditional, being exhibited only when expression of the gene is induced. This would allow propagation of the mutant in the absence of expression of the mutant phenotype, so that the inducible dominant negative mutations can be used in a similar way to temperature-sensitive mutations, for example to define intermediates and steps in some process or pathway. Furthermore, as the mutation is dominant and the phenotype is exhibited in

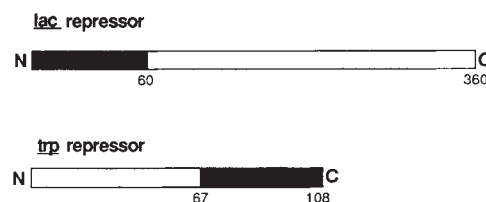


Fig. 4 Structure of *lac* repressor and *trp* repressor polypeptide chains. The structures of *lac* repressor (360 amino acids) and *trp* repressor (108 amino acids) monomers are shown (not drawn to physical scale). The filled regions contain the DNA binding domain, and the open regions contain the oligomerization domain.

the presence of functional wild-type genes, the consequences of inactivating the function of genes present in multiple copies can be studied without having to inactivate each copy of the gene.

How it might work

The efficacy of this approach depends on the likelihood of producing inhibitory polypeptides of the types described above. There are, in fact, a number of precedents for the kind of inhibitory interaction invoked here. In this section I shall explain how these precedents illustrate how the dominant negative mutation approach might work.

Gene-regulatory proteins. Naturally occurring dominant negative mutations affecting the multimeric repressor proteins for the *lac* and *trp* operons, and the lambda repressor have been identified, demonstrating that it is possible to inactivate the DNA-binding domain of such proteins without affecting the oligomerization domain. For example, many amino-acid substitutions in the DNA-binding domain of the *lac* repressor (Fig. 4) create subunits that inhibit wild-type repressor activity when overproduced (reviewed in ref. 27). The oligomerization domain of the *lac* repressor is apparently distinct from its DNA-binding domain, so that mixed aggregates containing mutant and wild-type subunits can be formed and are unable to bind DNA. A dominant negative version of the *trp* repressor can be produced simply by removing its carboxy-terminus²⁸ (Fig. 4), which generates a polypeptide of the form of derivative 1 (Fig. 3). Cloned genes whose *in vivo* function is not known might contain sequence motifs characteristic of DNA-binding proteins, for example, the 'helix-turn-helix' motif^{29,30}, which would provide a guide for tailoring gene derivatives of type 5.

There is a second way that inhibitory derivatives of DNA-binding proteins can be created: mutations could turn a transcriptional activator into a repressor. This could occur if the mutation left the DNA-binding domain intact, but destroyed the activity of the transcriptional activation domain. As an example of this, derivatives of the yeast GCN4 protein lacking the domain necessary for transcriptional activation block the function of the wild-type protein³¹. Similarly, mutations of the transcriptional activator encoded by the *x* gene of the human T-cell leukaemia virus HTLV-II have been identified that block the function of the wild-type protein³². It should be noted that it is not necessary in principle for a DNA-binding protein to work as a multimer for this strategy to be applicable—inhibitory versions of a protein that works as a monomer might be produced if they bound to the sites recognized by the wild-type protein but did not carry out the requisite activity.

Structures and machines. Proteins that make up structures such as muscle or the mitotic apparatus, or that are involved in cellular morphology appear exquisitely sensitive to defects in monomeric components. For example, mutations in the major actin gene of the nematode have dominant effects on movement³³. Microtubule function in *Drosophila* sperm is hypersensitive to the presence of defective subunits³⁴, as is body-shape in the nematode^{35,36}. The incorporation of inactive polypeptide subunits into a multiprotein aggregate such as the 'machines' thought to carry out DNA replication and recombination³⁷

would be expected to disrupt their function in a dominant manner.

Enzymes. Studies of hybrids formed between wild-type and mutant subunits of the enzyme aspartate transcarbamoylase (ATCase) of *Escherichia coli* illustrate two different ways in which mixed aggregates can have defects in enzyme function. Firstly, mutant subunits altered in active-site residues can associate with wild-type subunits to form non-functional aggregates³⁸, apparently because the active site of the enzyme is composed of residues from adjacent subunits³⁹. Secondly, subunits with changes outside the active site can also inhibit enzyme function in a dominant manner: mixed aggregates fail to respond to binding of a ligand apparently because the protein is 'frozen' in an inactive conformation⁴⁰. Enzymes with many subunits, such as glutamine synthetase with 12 (refs 41 and 42), would be expected to be particularly sensitive to mutant subunits.

Interpretation and inference

As with any genetic method, the use of dominant negative mutant derivatives to learn about gene and protein function requires interpretation and inference. Thus, although the approach is designed to disrupt a wild-type gene function, the overproduction of an inactive product might have the opposite effect—increased activity of the wild-type protein. This might occur if the defective form of the protein titrated a cellular inhibitor. An apparent example of this effect has recently been described for the yeast activator protein GAL4 (ref. 43). A fragment of the catalytic subunit of cyclic AMP-dependent protein kinase might similarly titrate its regulatory (inhibitory) subunit. The overproduction of different derivatives of a protein can thus have very different consequences, some inhibiting protein activity, others enhancing protein activity.

In some cases it is straightforward to determine whether overproduction of the mutant leads to increased or decreased activity of the cellular protein—if the protein in question is an enzyme, for example, in which case assaying enzyme activity would suffice. For a putative regulatory protein of unknown function, however, determining whether the derivative leads to increased or decreased activity of the cellular protein is not necessarily so straightforward, although it is crucial for inferring whether the wild-type protein is an activator or a repressor. Consider, for example, a protein that binds to the regulatory region of a liver-specific gene, and suppose that a mutant derivative of the protein causes hyperexpression of the liver-specific gene. This might mean that the wild-type protein is an activator

and that the mutant derivative enhances its activity, as described for GAL4 protein. Alternatively, the wild-type protein might be a repressor and the derivative might be acting as a classical dominant negative mutation. Deciding on which of these possible explanations is the case requires additional analysis.

Concluding remarks

The use of dominant negative mutations could provide information on the *in vivo* function of a diverse array of cloned genes and gene segments. The ability to manipulate genes *in vitro* makes it possible to design inhibitory products based on the principle that the activities of a protein can be separately mutated. The method has the genetic virtues of causing a conditional defect and being dominant, allowing the functional inactivation of redundant genes. Dominant negative mutations may provide a route to new information on protein-protein interactions and hence on designing new inhibitory polypeptides.

Finally, although the focus of this article has been methodological and, in a sense, artificial, the same type of complex mutations discussed in this article might be responsible for some cases of cellular transformation. Both dominant and recessive mutations are known to be able to cause cellular transformation. The dominant mutations in some cases result in hyperactivity of a wild-type product due to overexpression or modification of the product. Recessive mutations, such as that causing retinoblastoma, involve the loss of wild-type function. There is every reason to believe that loss of function could also result from dominant negative mutations. Production of a dominant negative oncogene *in vivo* could involve precisely the same types of events as those described for generating a dominant negative mutation by *in vitro* manipulations. One step would involve modifying the gene product itself to create a potential inhibitory polypeptide. The second step would be an event such as gene amplification that leads to overexpression of this mutant form. The question of whether some oncogenes have resulted from dominant negative mutations is not just a matter of genetic formalism: to understand how the product of an oncogene leads to uncontrolled cell growth, it is crucial to know whether the primary effect of the oncogene is to cause hyperactivity or inhibition of normal cellular activity.

I thank Bob Sauer and Harold Varmus, the members of my laboratory, and my colleagues at UCSF and elsewhere for valuable discussion and comments on the manuscript, as well as for directing me to relevant articles, and Paul Sternberg for the figures.

1. Rebagliati, M. R., Weeks, D. L., Harvey, R. P. & Melton, D. A. *Cell* **42**, 769–777 (1985).
2. Young, R. A. & Davis, R. W. *Proc. natn. Acad. Sci. U.S.A.* **80**, 1194–1198 (1983).
3. Muller, H. J. in *Proceedings of the Sixth International Congress of Genetics* (ed. Jones, D. F.) 213–255 (Brooklyn Botanic Gardens, Menasha, Wisconsin, 1932).
4. Scherer, S. & Davis, R. W. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4951–4955 (1979).
5. Rothstein, R. J. *Meth. Enzym.* **101**, 202–211 (1983).
6. Russell, P. & Nurse, P. *Cell* **45**, 145–153 (1986).
7. Miller, B. L., Miller, K. Y. & Timberlake, W. E. *Molec. cell. Biol.* **5**, 1714–1721 (1985).
8. May, G. S., Weatherbee, J. A., Gambino, J., Tsang, M. L.-S. & Morris, N. R. in *Molecular Genetics of Filamentous Fungi* (ed. Timberlake, W. E.) 239–251 (Liss, New York, 1985).
9. Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. *Nature* **317**, 230–234 (1985).
10. Thomas, K. R. & Capecchi, M. R. *Nature* **324**, 34–38 (1986).
11. Izant, J. G. & Weintraub, H. *Cell* **36**, 1007–1015 (1984).
12. Melton, D. A. *Proc. natn. Acad. Sci. U.S.A.* **82**, 144–148 (1985).
13. North, G. *Nature* **313**, 635 (1985).
14. Rosenberg, U. B., Preiss, A., Seifert, E., Jäckle, H. & Knipfle, D. C. *Nature* **313**, 703–706 (1985).
15. Crowley, T. E., Nellen, W., Gomer, R. H. & Firtel, R. A. *Cell* **43**, 633–641 (1985).
16. Nishikura, K. & Murray, J. M. *Molec. cell. Biol.* **7**, 639–649 (1987).
17. Kim, S. K. & Wold, B. J. *Cell* **42**, 129–138 (1985).
18. Rebagliati, M. R. & Melton, D. A. *Cell* **48**, 599–605 (1987).
19. Bass, B. L. & Weintraub, H. *Cell* **48**, 607–613 (1987).
20. Meeks-Wagner, D. & Hartwell, L. H. *Cell* **44**, 43–52 (1986).
21. Blose, S. H., Meltzer, D. I. & Feramisco, J. R. *J. Cell Biol.* **98**, 847–858 (1984).
22. Wehland, J. & Willingham, M. C. *J. Cell Biol.* **97**, 1476–1490 (1983).
23. Feramisco, J. R. *et al.* *Nature* **314**, 639–642 (1985).
24. Kung, H.-F., Smith, M. R., Bekesi, E., Manne, V. & Stacey, D. W. *Expl. Cell Res.* **162**, 363–371 (1986).
25. Chandler, V. L., Maler, B. A. & Yamamoto, K. R. *Cell* **33**, 489–499 (1983).
26. Swift, G. H., Hammer, R. E., MacDonald, R. J. & Brinster, R. L. *Cell* **38**, 639–646 (1984).
27. Miller, J. H. in *The Operon* (eds Miller, J. H. & Reznikoff, W. S.) 31–88 (Cold Spring Harbor Press, New York, 1978).
28. Kelley, R. L. & Yanofsky, C. *Proc. natn. Acad. Sci. U.S.A.* **82**, 483–487 (1985).
29. Matthews, B. W., Ohlendorf, D. H., Anderson, W. F. & Takeda, Y. *Proc. natn. Acad. Sci. U.S.A.* **79**, 1428–1432 (1982).
30. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M. & Pabo, C. O. *Nature* **298**, 447–451 (1982).
31. Hope, I. A. & Struhl, K. *Cell* **46**, 885–894 (1986).
32. Wachsmann, W. *et al.* *Science* **235**, 674–677 (1987).
33. Waterston, R. H., Hirsh, D. & Lane, T. R. *J. molec. Biol.* **180**, 473–496 (1984).
34. Fuller, M. T. in *Gametogenesis and the Early Embryo* (ed. Gall, J. G.) 19–41 (Liss, New York, 1986).
35. Kusch, M. & Edgar, R. S. *Genetics* **113**, 621–639 (1986).
36. Park, E. C. & Horvitz, H. R. *Genetics* **113**, 821–852 (1986).
37. Alberts, B. M. *Cold Spring Harb. Symp. quant. Biol.* **49**, 1–12 (1984).
38. Wente, S. R. & Schachman, H. K. *Proc. natn. Acad. Sci. U.S.A.* **84**, 31–34 (1987).
39. Robey, E. A. & Schachman, H. K. *Proc. natn. Acad. Sci. U.S.A.* **82**, 361–365 (1985).
40. Gibbons, I., Flatgaard, J. E. & Schachman, H. K. *Proc. natn. Acad. Sci. U.S.A.* **72**, 4298–4302 (1975).
41. Mitchell, A. P. *Genetics* **111**, 243–258 (1985).
42. Almasy, R. J., Janson, C. A., Hamlin, R., Xuong, N.-H. & Eisenberg, D. *Nature* **323**, 304–309 (1986).
43. Johnston, S. A., Zavortink, M. J., Debouck, C. & Hopper, J. E. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6553–6557 (1986).
44. De Zanne, A. & Spudich, J. A. *Science* **236**, 1086–1091 (1987).

Extragalactic radio sources with very large Faraday rotation

Tatsuji Kato*, Hiroto Tabara*, Makoto Inoue† & Ko Aizu‡

*Faculty of Education, Utsunomiya University, Mine, Utsunomiya 321, Japan

†Nobeyama Radio Observatory, Nobeyama, Minamisaku, Nagano 384-13, Japan

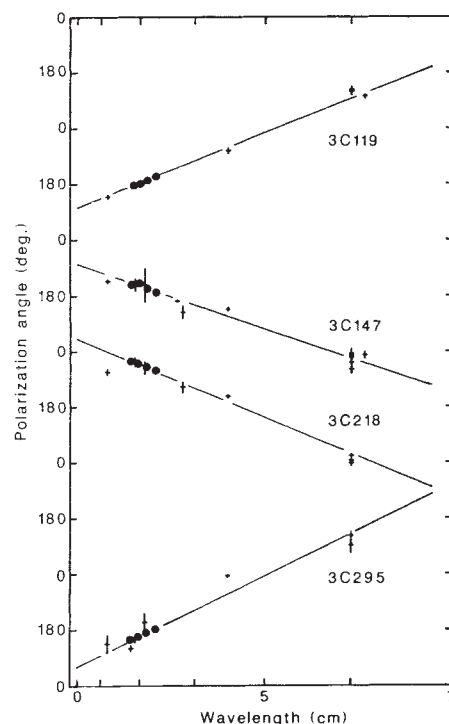
‡Katahira 2-24-3, Asao-Ku, Kawasaki 215, Japan

There are several radio sources with Faraday rotation measures larger than 500 rad m^{-2} (ref. 1). Because most of these sources are located near the galactic plane, the large rotation measures can be attributed to the galactic medium². However, a few sources at high galactic latitude also show large rotation, implying intrinsic Faraday rotation within the source. Until now, no sources have been discovered to have a rotation measure $>1,000 \text{ rad m}^{-2}$. Using a multichannel polarimeter at 10 GHz attached to the 45-m radio telescope at Nobeyama, we have observed about 100 extragalactic radio sources which were suspected to have large rotation from our polarization catalogue¹. Four of these 100 sources were found to have an intrinsic rotation measure $>1,000 \text{ rad m}^{-2}$. Here we report rotation measures which are two orders of magnitude larger than those observed so far for most extragalactic radio sources. Two of these four sources are classified as compact steep-spectrum sources.

In general, it is very difficult to determine a unique value for a rotation measure (RM) because of discrete sampling of observing wavelengths and the $n \times 180^\circ$ ambiguity of the polarization angle. Furthermore, if such a large Faraday rotation occurs within a source itself, large depolarization should occur³, and hence little polarization is observed at longer wavelengths. Thus polarimetry at higher frequencies with closely spaced wavelengths is essential to search for large RM sources. Using an updated version of the catalogue by Tabara and Inoue¹, we selected about 100 extragalactic radio sources which were suspected to have a large RM, based on a strong depolarization and/or a poor fit of the polarization angle to a square of wavelength (λ^2) law. From a statistical study of RM, Tabara and Inoue¹ set a limit of $|\text{RM}| \leq 1,000 \text{ rad m}^{-2}$ for measuring rotation, but this cannot exclude a few sources having RM larger than $1,000 \text{ rad m}^{-2}$. In fact, Inoue *et al.* discovered $\text{RM} = -1,660 \text{ rad m}^{-2}$ near the centre of our Galaxy⁴ and some sources such as Vir A and Cyg A are known to have an RM around $1,000 \text{ rad m}^{-2}$ (refs 5, 6).

The search for large RM was carried out with the 45-m telescope of the Nobeyama Radio Observatory on 3-4 January and 5-7 June 1985. The polarimeter consists of a half-wave plate capable of rotation and a polarization splitter. The difference of the orthogonal polarizations was obtained by switching the polarizations at 250 Hz. The receiver, with a typical system temperature of 130 K, has an instantaneous bandwidth of 2 GHz. At the intermediate frequency (IF) stage, four contiguous 500-MHz passband filters were set with sky frequencies centred at

Fig. 1 Polarization angle versus square of wavelength for large RM sources, from top to bottom 3C119, 3C147, 3C218, and 3C295, respectively. The solid line shows the best fit RM for each source. The values are $+1,728$, $-1,510$, $-1,850$, and $+2,250 \text{ rad m}^{-2}$ from top (3C119) to bottom (3C295), respectively. Filled circles around 3 cm show this observation. In the calculations of RMs, available data at wavelengths shorter than 6 cm in a catalogue by Tabara and Inoue¹ were used.



9.05, 9.55, 10.05 and 10.55 GHz. This enables us to derive larger RMs without the $n \times 180^\circ$ ambiguity in the polarization angle. The antenna half-power beamwidth was 2.7 arc min at 10 GHz. Instrumental polarizations were calibrated using 3C84 and DR21 at various elevations. The absolute polarization angle was calibrated by observing 3C286 and 3C348. By the repeatability of frequent observations of the calibrators, we estimated the accuracies of the polarization degree and the polarization angle to be $\leq 0.2\%$ and 0.5° , respectively, in all four channels at the elevation angles between 20° and 70° .

We discovered that four sources (3C119, 3C147, 3C218 and 3C295) have $|\text{RM}| > 1,000 \text{ rad m}^{-2}$. None of the other candidates shows clear evidence of $|\text{RM}| \geq 500 \text{ rad m}^{-2}$, and the results will appear elsewhere. Figure 1 shows the polarization angles versus λ^2 plot for these four sources together with catalogued data. Observations at $\lambda \leq 6 \text{ cm}$ are well fitted to straight lines, so that we can confirm our results. From observations at $\lambda \leq 3 \text{ cm}$ we derived the RM for 3C295 as $+1,740 \text{ rad m}^{-2}$, but the best fit RM is $+2,250 \text{ rad m}^{-2}$ with all observations at $\lambda \leq 6 \text{ cm}$ together with our observation. This may be explained by the fact that 3C295 is one of the powerful normal doubles and has hotspots like Cyg A⁷. Thus the component we observed may be different from that of longer wavelengths observations. The source parameters are summarized in Tables 1 and 2.

As these sources show remarkably large Faraday depolarization at centimetre wavelengths, it is likely that the origin of these large RM is due to plasma within the emitting region of the synchrotron radiation. If thermal electrons coexist with relativistic electrons, internal Faraday depolarization occurs³. In fact, all sources in Table 1 show depolarization between 3

Table 1 Source parameters

Source name	Identification	z	RM (rad m^{-2})	Intrinsic polar- ization angle (deg.)	Polarization degree (%)
					3 cm 6 cm
3C119 (0429+415)	QSO CSS	0.408	$+1,728 \pm 3$	97 ± 2	13.7 1.5
3C147 (0538+498)	QSO CSS	0.545	$-1,510 \pm 50$	111 ± 9	1.2 0.5
3C218 (0915-118)	cD	0.065	$-1,850 \pm 15$	52 ± 3	2.2 0.5
3C295 (1409+524)	G	0.461	$+2,250 \pm 35$	51 ± 3	1.6 0.2

Table 2 Polarization degree and angle of the four channels

Source name	Wavelength (cm)	Polarization degree (%)	Polarization angle (deg.)
3C119	2.844	15.4 ± 0.5	176.9 ± 1.0
	2.985	14.8 0.2	4.1 0.9
	3.141	13.2 0.2	15.3 1.0
	3.315	11.4 0.3	25.7 0.9
3C147	2.844	1.3 0.2	41.7 4.0
	2.985	1.5 0.2	44.8 4.1
	3.141	1.0 0.2	26.2 2.7
	3.315	0.8 0.2	10.8 2.7
3C218	2.844	2.4 0.2	148.5 2.4
	2.985	2.1 0.2	141.8 2.5
	3.141	2.1 0.2	130.0 2.4
	3.315	2.0 0.2	120.9 2.5
3C295	2.844	1.9 0.2	155.9 3.0
	2.985	1.8 0.2	159.8 3.7
	3.141	1.5 0.2	173.2 3.4
	3.315	1.2 0.2	3.8 3.8

and 6 cm. Even for sources near the galactic plane, such as 3C119 and 3C147, a remarkable depolarization can be seen (Fig. 2). The amount of RM originating in our Galaxy is about -10 rad m^{-2} for other lines of sight around these sources¹, which is roughly two orders of magnitude lower than our results.

Two (3C119 and 3C147) of the four sources are classified as compact steep-spectrum (CSS) sources⁸. They show a structure consisting of a core- or nucleus-like component together with (knotty) jet-like complex features embedded in an extended and/or diffuse component^{9,10}.

The large RM observed in CSS sources can be understood as being due to a dense medium which surrounds the nucleus^{11,12}. Jets coming out from the nucleus heavily interact with the environment, and hence irregularities in the medium and/or in the magnetic field result in depolarization³. In addition, thermal plasma in the interacting region (or emitting region) may also be responsible for the depolarization at longer wavelengths³.

Finally, as for 3C218 our recent observation at 22 GHz with the five-element interferometer at Nobeyama Radio Observatory shows that its structure is different from the other three: the inner part has 's-shaped' structure which is almost the same size as that of the optical galaxy (manuscript in preparation). The

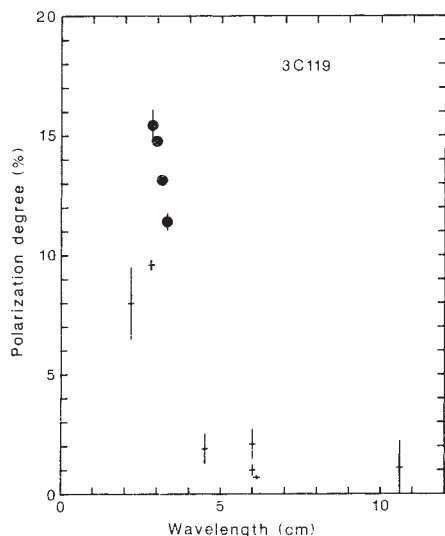


Fig. 2 Polarization degree versus wavelength for CSS quasar 3C119. Filled circles at around 3 cm indicate this observation. From shorter to longer wavelengths, polarization degrees are 15.4, 14.8, 13.2 and 11.4%, respectively. Remarkable Faraday depolarization can be seen at our 10 GHz band.

surrounding medium which would confine the jets may be, in this case, responsible for the large RM.

This work was carried out under the common observation programme at Nobeyama Radio Observatory (NRO). Nobeyama Radio Observatory, a branch of Tokyo Astronomical Observatory, University of Tokyo, is a facility open for general use by researchers in the field of astronomy and astrophysics.

Received 28 April; accepted 21 July 1987.

1. Tabara, H. & Inoue, M. *Astr. Astrophys. Suppl.* **39**, 379-393 (1980).
2. Inoue, M. & Tabara, H. *Publs astr. soc. Japan* **33**, 603-615 (1981).
3. Burn, B. J. *Mon. Not. R. astr. Soc.* **133**, 67-83 (1966).
4. Inoue, M. *et al. Publs astr. soc. Japan* **36**, 633-638 (1984).
5. Simard-Normandin, M. *et al. Astrophys. J. Suppl.* **45**, 97-111 (1981).
6. Dreher, J. W. *et al. Astrophys. J.* **316**, 611-625 (1987).
7. Wilkinson, P. N. *IAU Symp.* **97**, 149-156 (1982).
8. Pearson, T. J. *et al. Astr. J.* **90**, 738-755 (1985).
9. Fanti, C. *et al. Astr. Astrophys.* **170**, 10-19 (1986).
10. Preuss, E. *et al. IAU Symp.* **97**, 289-290 (1982).
11. van Breugel, W. J. M. *Astr. J.* **89**, 5-22 (1984).
12. Wilkinson, P. N. *et al. Nature* **308**, 619-621 (1984).

Radio-interferometric imaging of the subsurface emissions from the planet Mercury

J. O. Burns*, G. R. Gisler†, J. E. Borovsky†, D. N. Baker†‡ & M. Zeilik*

* Institute for Astrophysics, Department of Physics and Astronomy, The University of New Mexico, Albuquerque, New Mexico 87131, USA

† Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Mercury is the least studied of the terrestrial planets. Its interior and subsurface properties are still largely unknown. The subsurface can be probed at microwave frequencies where the electrical skin depth is $\sim 70 \text{ cm}$ at 5 GHz¹. Previous ground-based observations^{2,3} were unable to resolve details of Mercury's disk, and Mariner 10 did not carry a microwave radiometer. The Very Large Array (VLA) was used to make the first centimetre-wavelength interferometric observations that resolve the disk. We have mapped the distribution of total and polarized intensities from the planet's subsurface layers. We report the first detection of a hot pole along the hermean equator, which we model as black-body reradiation from preferential diurnal heating. These observations appear to rule out any internal sources of heat within Mercury. We also find polarized emission from the limb of the planet, which is understood in terms of the dielectric properties of the hermean surface.

VLA⁴ observations of Mercury at 5 GHz were performed on 6 July 1986. On this date, the planet had an eastern elongation of 22°, a phase of 0.22, a geocentric distance of 0.679 AU, and a diameter of 10 arc s. The phase centre of the array was pointed at the geometric centre of Mercury and this position was tracked for a period of 5 h. The resolution of these VLA observations was $\sim 1 \text{ arc s}$.

After calibration, the visibility data were Fourier transformed into a 'dirty' map and the map was 'cleaned' of sidelobes using the NRAO AIPS software^{5,6}. A cleaned map of Mercury is shown in Fig. 1. Flux densities were converted into brightness temperatures by the relation $S = 2kTv^2\Omega/c^2$ (Rayleigh-Jeans approximation for the Planck black body), where Ω is the solid angle of the clean beam. Note the presence of a region of slightly higher temperature in the right-hand, central portion of the map. We believe this is the 'hot pole' previously predicted in models⁷ but never before observed. We also performed a maximum

‡ Present address: NASA, Goddard Space Flight Center, Code 690 Greenbelt, Maryland 20771, USA.

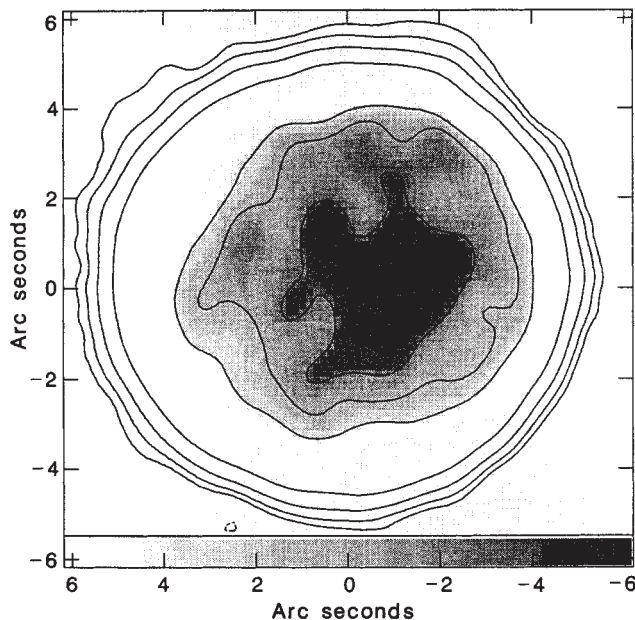


Fig. 1 VLA 5-GHz map of the surface temperature distribution on Mercury. This map has been convolved with a circular gaussian beam of 1 arc s FWHM, producing about 10 resolution elements across the planetary disk. The grey-scale pattern ranges from 300 K (lightest) to a saturation above 360 K. Contour levels are 20 K, 40 K, 80 K, 160 K, 310 K, 330 K, 350 K and 370 K. Note the presence of a 'hot pole'.

entropy method⁸ of deconvolution of the dirty map which produced comparable structures and intensities as the cleaned map, including the hot pole.

Although we believe the gross structures in Fig. 1 are real and reproducible, the finer-scale mottling on the map may be produced by incomplete sampling of the visibility function at high spatial frequencies. The r.m.s. noise on the map is ~ 3.5 K. This should be used cautiously as an error in relative temperature measurement as the mottling along the disk of the planet may be viewed as a potential error. The error in the absolute temperature also contains a calibration term that we conservatively estimate to be 5%. Thus, the peak temperature on the map is 378 ± 19 K, in good agreement with previous measurements.

The peculiar circumstances of Mercury, namely a highly eccentric orbit and a phase locking, in which three sidereal rotations are completed in two complete orbits, result in an unusual heating of the planet's surface by the Sun⁷. The local solar day endures for two full orbits, or ~ 176 Earth days; perihelion passage occurs at noon for hermographic longitudes 0° and 180° . Because the planet moves fastest in its orbit near perihelion, while its rotation rate is constant, the Sun appears to hover overhead and indeed executes a small retrograde loop (see left inset in Fig. 2). The net effect is that portions of the surface near 0° and 180° longitude are heated much more strongly by the Sun than the rest of the surface.

A complete knowledge of the thermal, electrical and radiative properties of the hermean regolith is required to calculate the temperature at depth from such a thermal history as a function of longitude and latitude. Such calculations have been done in the past (refs 2, 3, 9 and R. Landau, personal communication) to aid in the interpretation of single-dish and low-resolution interferometric radio observations of the planet. The published results generally consist of integrated disk temperatures as a function of observational phase angle and sub-Earth longitude. Our VLA data require knowledge of the temperature at depth for many resolved elements across the disk—information not extractable from the published results of previous thermal calculations. We intend to do such calculations in the future, when

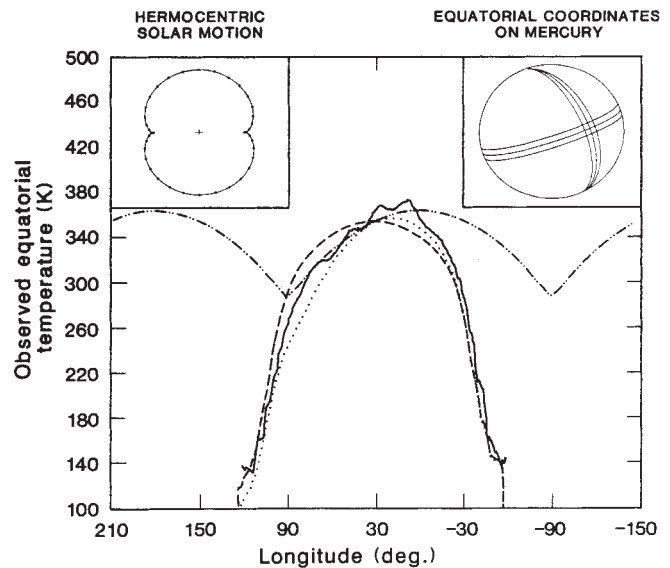


Fig. 2 The left inset illustrates the apparent motion of the Sun in a reference frame centred on Mercury, and rotating with Mercury's sidereal rotation rate. Hermographic longitude zero is at three o'clock. The tick marks on the orbit are separated by four days. The right inset illustrates the aspect presented by Mercury at the epoch of observation. The equator, circles of latitude $\pm 5^\circ$, and meridians of longitude at $0, \pm 5^\circ$ are plotted. The intersection of these lines should be the hottest region on the planet's surface. In the main figure, the dot-dash line is the equilibrium subsurface temperature on the equator calculated as a function of longitude from the total insolation received at that longitude over a hermean day. The dashed line is the brightness temperature which would be observed from a uniform temperature sphere, convolved with a gaussian beam to match our observations. The dotted line is the equatorial brightness temperature which would be observed from a sphere whose equatorial temperature profile is given by the dot-dash line. The solid line presents our observations as a profile across the map along the equator.

we have complete VLA and infrared data. For now, we offer a much simpler calculation that is independent of the nature of the regolith.

For simplicity, we confine our attention to the equator. Knowledge of the orbit of Mercury about the Sun immediately gives us the insolation pattern for any point on the planet. For points equally spaced along the equator, we calculate the total solar energy received by the surface over the course of a complete hermean day. The energy received is balanced by the energy radiated by the surface; we do this calculation separately for each longitude to yield an equilibrium subsurface temperature as a function of longitude. We present the results of this calculation for a sphere of zero albedo and unit emissivity in Fig. 2. The theoretical curve is only slightly depressed if we use an albedo of 5% and an emissivity of 95%.

At the epoch of observation, the hermographic longitude and latitude at the sub-Earth point are $+30.92^\circ$ and $+8.92^\circ$, so that Mercury presents to us the aspect shown in the right inset in Fig. 2. The hottest part of the planet's surface is the intersection of the prime meridian with the equator. It is not, however, the hottest part of the observed map because the flux of radiation received at a distant point from a surface element on a perfect sphere is diminished by the cosine of the angle the surface normally makes with the line to the Earth. Converting flux to temperature units, this diminution factor becomes $(\cos \theta_0)^{1/4}$. Using this diminution factor, and convolving with the gaussian beam of the VLA, the model gives the dotted line in Fig. 2.

The observations follow the model surprisingly well, considering the model's utter simplicity, except for the quasi-periodic structures (possible artefacts of image reconstruction). Our

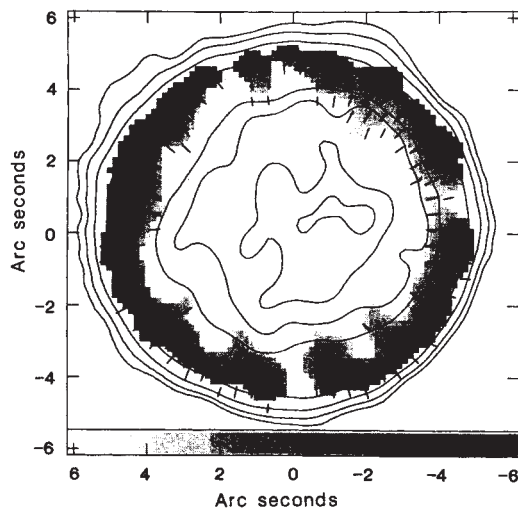


Fig. 3 Grey-scale representation of the percentage linear polarization at 5 GHz overlaid onto a contour map of the total intensity. Specifically, the percentage polarization is defined by $100\% \times (Q^2 + U^2)^{1/2} / I$ where I is the total intensity, and Q and U are Stokes polarization parameters. The grey-scale ranges from 2% (lightest) to 12% (darkest). Contour levels are the same as in Fig. 1. Note the edge-brightened appearance of the polarized emission. The dashed lines are polarization electric field vectors.

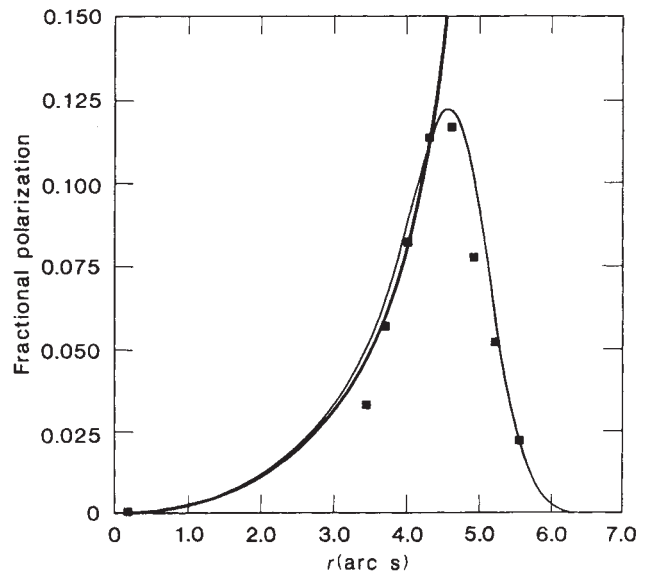


Fig. 4 Radial distribution of fractional linear polarization compared with model calculations. The filled squares are the average observed data points. The thick curve is the smooth spherical model prediction for a dielectric constant of 1.55 discussed in the text for infinite resolution. The thin curve is the model convolved with the 1 arc s FWHM beam of the VLA. Surface roughness effects are expected to depolarize the observed emission. Therefore, this simple model probably produced a lower limit to the surface dielectric constant of Mercury.

observations can immediately rule out thermal models of Mercury which have an internal heat source producing in excess of 10–20 K at the surface.

Although the radiation emerging from the subsurface layers of Mercury is thermal, and thus initially randomly polarized, the change in refractive index between the hermean surface and atmosphere will result in a net linear polarization of the refracted radio waves. Therefore, by mapping the distribution of percentage polarization of the radio emission, we can investigate the electrical and geometrical properties of the planet's surface. In Fig. 3, a grey-scale representation of the percentage linear polarization is overlaid onto the total intensity contour map. A clear ring of polarization is seen toward the limb of the planet. The polarization ranges from zero at the centre of Mercury to a peak of about 12% at the limb (see also Fig. 4). The dashed lines in Fig. 3 illustrate both the magnitude and direction of the electric field vectors of the radiation. Note that these E-vectors all point radially in toward the centre of Mercury.

Assuming Mercury to be a smooth spherical dielectric, the ratio of the azimuthal to radial electrical field components (as projected on the observed disk) is predicted¹⁰ to be

$$\frac{E_a}{E_r} = \frac{\epsilon + \cos \theta_0 \left[1 - \frac{\sin^2 \theta_0}{\epsilon_2} \right]^{-1/2}}{1 + \epsilon \cos \theta_0 \left[1 - \frac{\sin^2 \theta_0}{\epsilon^2} \right]^{-1/2}}$$

where ϵ is the dielectric constant of the surface of Mercury at radio frequencies, and θ_0 is the angle between the surface normal and the line of sight to Earth. At the sub-Earth point (where the radio rays to the observer are normal to the planet's surface), the emission is unpolarized. At the planetary limb, the polarization is maximum with $E_r/E_a = \epsilon$. The fractional polarization $(E_r^2 - E_a^2)/(E_r^2 + E_a^2)$ is plotted as the thick curve in Fig. 4 as a function of the radial distance from the centre of the disk of Mercury. This curve is then convolved with a gaussian radio

beam that has a full-width at half maximum of 1 arc s (the thin curve in Fig. 4). The observed distribution of fractional polarization at 6 cm is shown as the solid squares. The polarization data is consistent with Mercury being a smooth sphere with a surface dielectric constant $\epsilon = 1.55$. If, as expected, the surface of Mercury is not smooth (on the scale of the 6-cm radio wavelength), then the observed emission is depolarized with respect to a smooth sphere¹¹. Therefore, our derived value of ϵ should be viewed as a lower limit.

The distribution of total and polarized microwave emissions from Mercury are well explained by simple models using black-body radiation from the subsurface layers and the refractive properties of the surface. The success of these first, preliminary VLA observations and the models leads us to believe that we can further improve upon the known thermal and electrical properties of the hermean regolith by collecting more accurate, higher resolution radio data. These observations are now under way.

We are indebted to Skip Newhall of JPL for providing us with an excellent ephemeris of Mercury, and to R. Ekers and B. Clark scheduling the VLA observations. The National Radio Astronomy Observation is operated by Associated Universities, Inc., under NSF contract. Work at Los Alamos was supported by the US Department of Energy.

Received 8 June; accepted 22 July 1987.

1. Baker, D. N., Borovsky, J. E., Burns, J. O., Gisler, G. R. & Zeilik, M. *J. geophys. Res.* **92**, 4707–4712 (1987).
2. Cuzzi, J. N. *Astrophys. J.* **189**, 577–586 (1974).
3. Golden, L. M. thesis, Univ. California, Berkeley (1977).
4. Thompson, A. R., Clark, B. G., Wade, C. M. & Napier, P. J. *Astrophys. J. Suppl. Ser.* **44**, 151–167 (1980).
5. *Astronomical Image Processing System (AIPS)* (National Radio Astronomy Observatory, Green Bank, West Virginia, 1986).
6. Clark, B. G. *Astr. Astrophys.* **89**, 377–378 (1980).
7. Morrison, D. *Space Sci. Rev.* **11**, 271–307 (1970).
8. Cornwell, T. J. & Evans, K. F. *Astr. Astrophys.* **143**, 77–83 (1985).
9. Morrison, D. *Smithsonian astrophys. Obs., Spec. Rep. No. 292* (1969).
10. Jackson, J. D. *Classical Electrodynamics* 2nd edn section 7.3 (Wiley, New York, 1975).
11. Golden, L. M. *Icarus* **38**, 451–455 (1979).

Oxygen ordering and superconductivity in $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$

C. U. Segre^{*†}, B. Dabrowski[‡], D. G. Hinks[‡],
K. Zhang^{*†}, J. D. Jorgensen[‡], M. A. Beno[‡]
& Ivan K. Schuller[‡]

^{*} Department of Physics, Illinois Institute of Technology, Chicago, Illinois 60616, USA

[‡] Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA

The observation of superconductivity above 90 K in the Y–Ba–Cu–O system^{1,2} has prompted a search for other structures that display high-temperature superconductivity. The discovery of superconductivity above 70 K in $\text{La}_{3-x}\text{Ba}_{3+x}\text{Cu}_6\text{O}_{14+y}$ ³ raises some questions concerning its relationship to the $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ structure and the origins of the high-temperature superconductivity. The original X-ray diffraction study of $\text{La}_3\text{Ba}_3\text{Cu}_6\text{O}_{14+y}$ indicated a tetragonal structure^{4,5}, on the basis of which it has been proposed that $\text{La}_{3-x}\text{Ba}_{3+x}\text{Cu}_6\text{O}_{14+y}$ is a new class of high-temperature superconductor which, in contrast to $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$, contains no Cu–O chains³. More recently, a neutron diffraction study of non-superconducting $\text{La}_3\text{Ba}_3\text{Cu}_6\text{O}_{14+y}$ ⁶ has indicated that the correct structure of this compound is simply the tetragonal $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ structure with partial substitution of lanthanum on the barium site^{7–12}. In the present neutron powder diffraction study of a series of $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ compounds, we show that the superconducting phase is a similarly disordered isomorph of the orthorhombic $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ structure^{13–19}. The region of solid solubility for this compound extends from $x = 0.5$ (La:Ba:Cu ratios of 3:3:6) to $x \approx 0.25$ and does not include the stoichiometric compound $x = 0$ ($\text{LaBa}_2\text{Cu}_3\text{O}_{7-\delta}$). The presence of Cu–O chains, and not just a formal copper charge state of >2 , is vital for the occurrence of superconductivity.

Five samples in the series $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ were prepared by mixing and grinding La_2O_3 , BaCO_3 and CuO powders in the correct metal stoichiometries and firing at 975 °C in flowing oxygen. The samples were then reground and refired repeatedly to improve homogeneity. A final annealing of all samples at 600 °C in flowing oxygen was necessary to ensure the presence of superconductivity. Neutron powder diffraction data were obtained using the Special Environment Powder Diffractometer (SEPD) at the Argonne National Laboratory Intense Pulsed Neutron Source (IPNS). Resistivities were measured in a closed-cycle refrigerator from 300 to 8.5 K using silicon-diode thermometry, by the standard four-probe technique. The neutron diffraction data for the five samples indicated that the $x = 0.5$, 0.375 and 0.25 samples were single phase but the two samples with the largest amount of barium, $x = 0.125$ and 0.0, had significant amounts of BaCuO_2 . Michel *et al.* have reported the synthesis of the stoichiometric $\text{LaBa}_2\text{Cu}_3\text{O}_{7+\delta}$ compound with

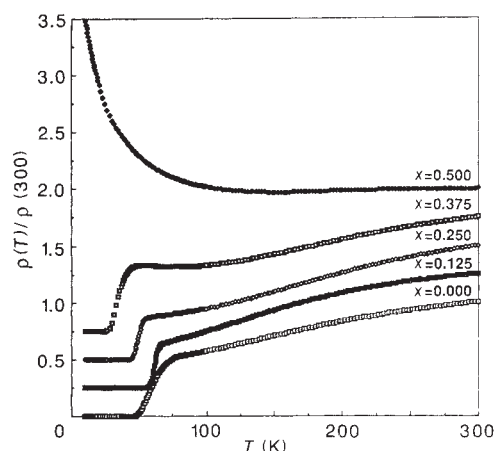


Fig. 1 Relative resistivity versus temperature for the five $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ samples ($x = 0, 0.125, 0.25, 0.375, 0.5$). The curves are displaced in the vertical direction for clarity.

$T_c = 75$ K (ref. 20), but their preparation technique is somewhat different and they give no details as to sample quality.

Figure 1 shows the relative resistivities as a function of temperature for the five $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ samples. At $x = 0.5$ the resistivity shows semiconducting behaviour, but as x is decreased, the samples become more metallic and superconductivity appears, with the transition temperature T_c rising to a maximum of ~ 60 K for the two samples $x = 0.125$ and $x = 0.0$. This behaviour is entirely consistent with previous work³, and indicates a solid-solution limit between $x = 0.25$ and $x = 0.125$. The broad superconducting transitions of even the single-phase samples are typical of solid solutions, where the stoichiometry is never microscopically identical throughout the sample.

The first-reported $\text{La}_3\text{Ba}_3\text{Cu}_6\text{O}_{14+y}$ structure^{4,5} is related, as far as metal atoms are concerned, to the tetragonal $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ structure simply by a doubling of the unit cell along the basal-plane diagonals. To test the recent neutron diffraction result⁶, the initial structural refinement of the non-superconducting $x = 0.5$ sample was done using both models (both with space group $P4/mmm$). In agreement with ref. 6, it was discovered that the smaller $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ tetragonal structure gave a better fit to the data. The remaining samples, all of which are superconducting, were found to be orthorhombic. The two single-phase samples, $x = 0.375$ and 0.25 , were fitted by using the orthorhombic $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ model^{13–19}; the remaining two samples, $x = 0.125$ and 0.0 , required the inclusion of BaCuO_2 as a second phase.

Because of the close relationship between the structure of the orthorhombic superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ and the tetragonal but non-superconducting $\text{La}(\text{Ba}_{1.50}\text{La}_{0.50})\text{Cu}_3\text{O}_{7+\delta}$ compound, both orthorhombic and tetragonal fits were attempted on all of the $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ samples. The $x = 0.50$ sample would only converge for the tetragonal $P4/mmm$ model. For the other four samples, however, both tetragonal and orthorhombic ($Pmmm$) models were found to converge. To determine which of the two models was correct, a statistical analysis of the refined

[†] Present address: Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA.

Table 1 Physical parameters of the $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ samples

x	$7+\delta$	n_{01}	n_{05}	a (Å)	b (Å)	c (Å)	T_c^* (K)	Cu charge
0.5	7.22(3)		0.61(3)		3.9112(1)	11.6975(3)	—	2.31(2)
0.375	7.25(7)	0.74(5)	0.51(5)	3.9090(1)	3.9152(1)	11.7322(3)	27–47	2.38(5)
0.25	7.18(5)	0.76(4)	0.42(3)	3.9112(1)	3.9181(1)	11.7534(3)	44–54	2.37(3)
0.25(0.125)†	7.11(6)	0.85(5)	0.26(3)	3.9085(1)	3.9286(2)	11.7624(5)	57–68	2.32(4)
0.25(0.0)†	7.25(5)	0.86(5)	0.39(3)	3.9137(1)	3.9238(1)	11.7550(4)	47–80	2.42(3)

* T_c data are listed as the interval $T^{R=0} - T_{\text{onset}}$.

† Numbers in parentheses refer to the starting compositions, which are also used as sample labels in the text. The first number refers to the stoichiometry actually used in the refinement (see text).

Error in last digit is shown in parentheses.

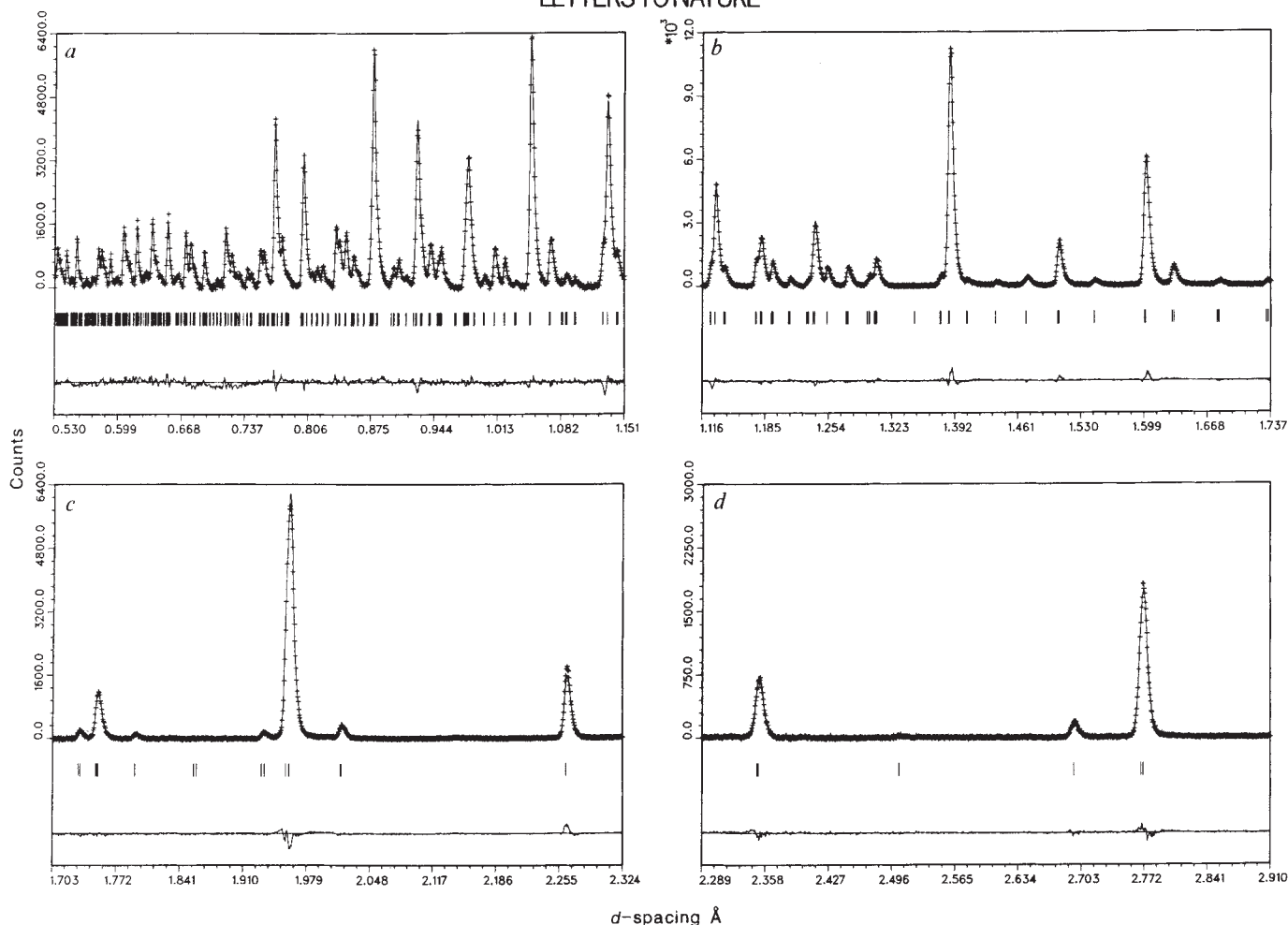


Fig. 2 *a-d*, Neutron diffraction spectrum for $\text{La}(\text{Ba}_{1.75}\text{La}_{0.25})\text{Cu}_3\text{O}_{7+\delta}$. The refined orthorhombic fit is shown as a line through the data points while the difference plot is shown in the lower portion of each frame.

data was required. First, the orthorhombic model was refined, then a second refinement was performed with constraints imposed to force the model to be tetragonal. The weighted profile R values for the two fits were then compared by the R -factor ratio test²¹. For the $x = 0.25$ sample, the orthorhombic fit had 33 variable parameters and weighted profile $R_0 = 5.527$ while the tetragonally constrained fit had 26 variable parameters and $R_1 = 5.910$. The ratio $R_1/R_0 = 1.069$ for the $\sim 3,500$ degrees of freedom and order $33 - 26 = 7$ indicates that the tetragonally constrained model can be rejected at the 99.5% confidence level. Figure 2 shows the neutron diffraction spectrum, the refined orthorhombic fit and the difference plot for the $x = 0.25$ sample. It should be noted that the orthorhombic splitting is not visually observable and only the high resolution and the broad range of d -spacings obtainable at IPNS allow us to detect the small orthorhombic distortion.

Figure 3 shows the general structure of the orthorhombic $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ compounds. In all of the orthorhombic samples both the O1 and O5 sites are partially occupied, with more oxygen atoms in the O1 site. The tetragonal sample has essentially the same structure but with the O1 and O5 sites equally occupied. Table 1 shows the refined oxygen stoichiometry, lattice parameters, T_c , formal copper charge state and O1 and O5 site occupancies for each of the samples investigated. The La, Cu1, Cu2, O2, O3 and O4 sites remained fully occupied when permitted to vary, so in the final refinement all atomic sites were held fully occupied except for the O1 and O5 sites in the basal plane. The ratio of barium to lanthanum on the Ba/La site was held fixed because, when allowed to vary, it refined to values not consistent with the starting composition in the single-phase samples. This instability in the Ba/La site

occupancy can be accounted for by realizing that when lanthanum, a much smaller atom, substitutes for barium, it is unlikely that it will occupy exactly the same spatial position or surroundings. As a consequence the thermal parameter obtained from the refinements is large and decreases as the site becomes more fully occupied with barium.

As mentioned above, the two samples with the highest superconducting transition temperature contain noticeable amounts of BaCuO_2 . This indicates that, for our preparation conditions, there is a minimum x for which the structure is stable. Because of the difficulty in refining the occupancy of the Ba/La site, it was not possible to extract the actual value of x for this endpoint. Instead, we chose to fix the composition at the lowest value of x for which we observed a single phase sample, $x = 0.25$. Although this is not the correct composition, changing this number had little effect on the trends observed in the important parameters of the resultant refined structure, that is, the formal charge state of copper and the ordering and occupancies of the O1 and O5 sites.

An important feature of these results is that the formal charge state of copper is always approximately +2.3, even in the non-superconducting sample ($x = 0.5$). This indicates that the copper charge state alone is not sufficient to produce superconductivity. Instead, only those samples that are orthorhombic, owing to a finite amount of oxygen-vacancy ordering, exhibit superconductivity. Figure 4 shows the zero-resistance temperature ($T^{R=0}$) plotted against an order parameter defined as the normalized difference in the O1 and O5 site occupancies, $\xi = (n_{\text{O1}} - n_{\text{O5}})/(n_{\text{O1}} + n_{\text{O5}})$. The points plotted in Fig. 4 fall on a smooth, monotonically varying curve and support our original suggestion that T_c is intimately connected to the presence of

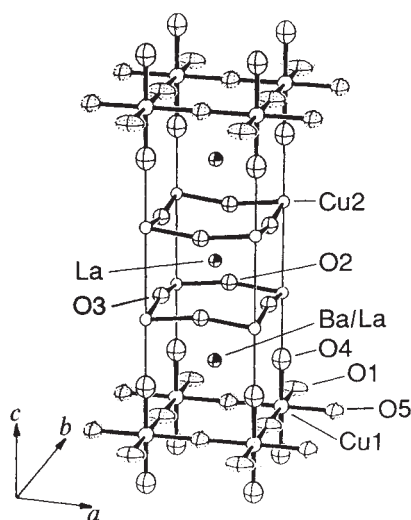


Fig. 3 General structure of the orthorhombic $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ compounds. The O1 and O5 sites are both partially occupied, with higher oxygen content in the O1 site.

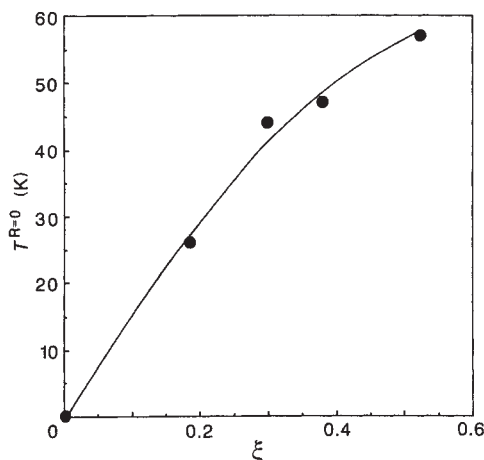


Fig. 4 Zero-resistance temperature ($T^{R=0}$) versus order parameter ($\xi = (n_{O1} - n_{O5}) / (n_{O1} + n_{O5})$) for the $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ compounds. The curve is merely a guide to the eye.

ordered Cu-O chains⁷. One would expect that if it were possible to synthesize a perfectly ordered $\text{LaBa}_2\text{Cu}_3\text{O}_{7-\delta}$ compound, it would have $\xi = 1$ and $T_c \approx 90$ K, as in orthorhombic $\text{RBa}_2\text{Cu}_3\text{O}_{7-\delta}$ ($\text{R} = \text{Y, Sm, Eu, Gd, Dy, Ho, Tm, Yb, Lu}$; $0 < \delta < 0.2$). Note that the resistivity data shown in Fig. 1 for the two-phase samples are not identical, as would be expected if both samples contained a majority phase with the same stoichiometry. The $x = 0.125$ sample has an appreciably higher $T^{R=0}$ than the $x = 0.0$ sample. The reason for this is that the neutron refinements for these two samples (which are a measure of the structural properties of the majority phase) imply that the $x = 0.125$ sample has a higher value of ξ , and consequently higher $T^{R=0}$ (which is also characteristic of the superconducting properties of the majority phase). Because of their two-phase character, residual inhomogeneities in these two samples could easily account for these effects. This analysis also points out the need for detailed determination of the oxygen stoichiometries and ordering for the understanding of the superconducting behaviour of metal oxides.

We have shown that the superconducting phase in the $\text{La}(\text{Ba}_{2-x}\text{La}_x)\text{Cu}_3\text{O}_{7+\delta}$ system is isostructural to the orthorhombic $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ and contains ordered Cu-O chains. This system is unique in that for a constant copper charge state of >2 , T_c is varied merely by changing the degree of order of the Cu-O chains. This contrasts with $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$, in which decreasing T_c is correlated with a decrease in copper oxidation as oxygen is removed (see, for example, ref. 22). Our data show that

ordering of the Cu-O chains is essential for superconductivity in the $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ structure.

This work is supported by Basic Energy Sciences, Division of Materials Sciences.

Note added in proof: The authors have recently become aware of the successful synthesis of orthorhombic $\text{LaBa}_2\text{Cu}_3\text{O}_{7+\delta}$ with $T_c \sim 90$ K (ref. 23).

Received 6 August; accepted 1 September 1987.

1. Wu, M. K. *et al.* *Phys. Rev. Lett.* **58**, 908 (1987).
2. Chu, C. W. *et al.* *Phys. Rev. Lett.* **58**, 405 (1987).
3. Mitzi, D. B. *et al.* (submitted).
4. Er-Rakho, L. *et al.* *J. Solid St. Chem.* **37**, 151 (1981).
5. Provost, J. *et al.* *Synth. Met.* **4**, 147 (1981).
6. David, W. I. F. *et al.* *Nature* **328**, 328-329 (1987).
7. Schuller, I. K. *et al.* *Solid St. Commun.* **63**, 385 (1987).
8. Jorgensen, J. D. *et al.* *Phys. Rev. B* (in the press).
9. Van Tendeloo, C. *et al.* *Solid St. Commun.* (in the press).
10. Santoro, A. *et al.* *Mater. Res. Bull.* **22**, 1007 (1987).
11. Katano, S. *et al.* *Jap. J. appl. Phys.* **26**, L1049 (1987).
12. Hewat, A. W. *et al.* (submitted).
13. Beno, M. A. *et al.* *Appl. Phys. Lett.* **51**, 57 (1987).
14. Greedan, J. D. *et al.* *Phys. Rev. B* **35**, 8770 (1987).
15. Capponi, J. J. *et al.* *Europhys. Lett.* **3**, 1301 (1987).
16. David, W. I. F. *et al.* *Nature* **327**, 310-312 (1987).
17. Beech, F. *et al.* *Phys. Rev. B* **35**, 8778 (1987).
18. Katano, S. *et al.* *Jap. J. appl. Phys.* **26**, L1046 (1987).
19. Cox, D. E. *et al.* *J. Phys. Chem. Solids* (in the press).
20. Michel, C. *et al.* *C. r. heb. Séanc. Acad. Sci., Paris* **304**, 1169 (1987).
21. *International Tables for X-ray Crystallography*, IV, 292 (Kynoch, Birmingham, 1974).
22. Jorgensen, J. D. *et al.* *Phys. Rev. Lett.* (submitted).
23. Maeda, A. *et al.* *Jap. J. appl. Phys.* **26**, L1366 (1987).

Microstructure and critical current of superconducting $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$

R. A. Camps, J. E. Evetts, B. A. Glowacki*,
S. B. Newcomb, R. E. Somekh & W. M. Stobbs

Department of Materials Science and Metallurgy,
University of Cambridge, Pembroke Street,
Cambridge CB2 3QZ, UK

In the few months that have passed since the discovery of high-transition-temperature superconducting oxides¹⁻³, remarkable progress has been made in identifying the structure of superconducting phases, improving our understanding of factors that control their synthesis, and determining their basic properties. In one area, however, progress has been disappointingly slow; that is, in the control and optimization of the superconducting critical current density, a property vital for effective technological application of these materials. Bulk samples display a critical current that is generally very low and, furthermore, drops steeply in an applied magnetic field. There is persuasive evidence, however, that low critical currents are not an inherent characteristic of the material^{4,5}, and recent work on thin epitaxial films⁶ has indicated that values as high as 10^6 A cm^{-2} can be achieved. It is also generally established that the critical current that can be carried by any superconductor depends predominantly on the local microstructure, its degree of uniformity and the nature of its characteristic defect structures⁷. We present here the main results of a detailed electron microscopical study of sintered $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ ⁸, and discuss their relevance to observed critical current and its modification by increased silicon impurity levels⁹. In the light of these results we assess the prospects for controlling microstructure development so as to optimize the critical current for technological application.

The agreed basic structure of the orthorhombic superconducting form of $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ is that of an oxygen-deficient perovskite, with Ba and Y in the sequence Ba-Y-Ba on the A sites of the $\text{A}_2\text{B}_2\text{O}_6$ structure, no oxygen on the Y planes and oxygen vacancies leaving tunnels in the [010] direction on the Cu planes

* Permanent address: Technical University of Wroclaw, Poland.

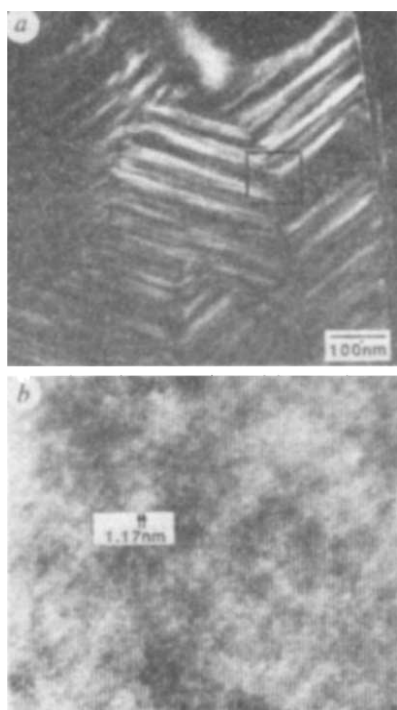


Fig. 1 *a*, Bright-field micrograph ($g=100$) showing different $\{113\}$ -habit domain-boundary regions in a single grain. The boundaries separate domains with a common c axis but alternating near-parallel b and a axes. *b*, Enlarged region of *a*, as boxed, with $g=003$, showing the lack of disruption of the c direction in the stress relaxation structure.

between the two neighbouring Ba planes^{10,11}. Above ~ 675 K the oxide undergoes a transformation to a tetragonal structure, with disordering of the oxygen vacancies; the relationship of the oxygen deficit, x , to the superconducting transition temperature T_c and the ratio of lattice constants b/a has recently been characterized^{12,13}. T_c is found to be strongly dependent on the precise value of the oxygen deficit in the range $x=0$ to 0.5 ¹²; in any structural assessment, therefore, it is important to concentrate on relationships between the oxygen ordering, the value of x and the ordering transformation.

Previously reported electron microscopical work^{14–16} has placed emphasis on the use of high-resolution techniques in an attempt to clarify the degree of uniformity of the $[010]$ oxygen vacancy tunnels. Such work is unlikely to be successful using any conventional approach because of the large anisotropic temperature factor for longitudinal observations in the x direction of the O atoms in the $[010]$ O–Cu–O inter-Ba chains (see, for example, ref. 11). Data on microstructure and characteristic defect structures is sparse and somewhat in conflict, and it is important to establish whether any of the observed features are artefacts of particular sample preparation methods. Ourmazd *et al.*¹⁴, working on ion-beam-thinned material, observed a high density of basal-plane faults which they interpreted as the response of the structure to changes in oxygen content. On the other hand, Hewat *et al.*¹⁵ did not observe a corresponding density of this type of faults in specimens prepared conventionally by grinding a suspension of crystals in dichloromethane. Furthermore, Hyde *et al.*¹⁶ noted that well-ordered material can apparently degrade in air at room temperature to exhibit basal-plane faulting. We have examined both dry-ground and argon-ion-beamed-thinned specimens and can confirm that the material as ground is well ordered and does not exhibit extensive basal-plane faulting, whereas ion-beam thinning can result in specimen faulting of the type described by Ourmazd *et al.*¹⁴.

The material investigated was prepared from a mixture of BaCO_3 , Y_2O_3 and CuO powders (Johnson Matthey, $>99.999\%$

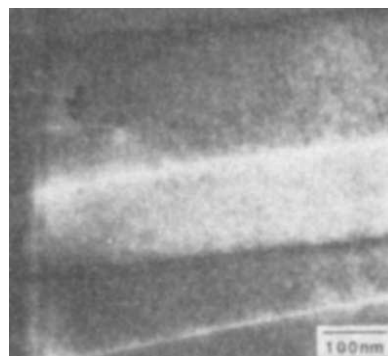


Fig. 2 Dark-field micrograph showing domain structure similar to that in Fig. 1*a* but with boundaries of approximately $\{100\}$ habit. The formation of a small arrowhead-shaped domain near the centre of a larger domain can be seen at the grain boundary (arrow).

purity) in stoichiometric proportion according to the formula $\text{YBa}_2\text{Cu}_3\text{O}_x$. The powders were dried, ground together in an agate ball mill, pressed into pellet form and heated to 900°C for 23 h in air. The product was re-ground and pressed, then heated to 950°C and cooled slowly to 200°C under oxygen. The material exhibited a T_c midpoint of 91.5 K, with a superconducting transition breadth of 1 K. X-ray diffractometer data indicated that the material was orthorhombic ($a=3.82$ Å, $b=3.88$ Å, $c=11.68$ Å). Ion-beam-thinned and dry-ground specimens were examined using standard bright- and dark-field techniques, convergent-beam electron diffraction, and energy-dispersive X-ray analysis at 120 kV in a Philips EM400T electron microscope. The oxide prepared as described exhibited a fine equiaxed grain structure with a grain size of 2 – 5 μm . Particularly noteworthy features of the microstructure are the almost universal occurrence of domain-type substructures within the grains, and the common occurrence at grain boundaries of narrow bands of structurally modified material and second-phase precipitates.

The morphology of the domain structure is typical of a stress relaxation as associated with a structural transformation, and in Fig. 1*a* the central part of the grain has parallel domains separated by boundaries of one $\{113\}$ habit, the peripheral domains being separated by boundaries of another $\{113\}$ habit. We suggest that this substructure is formed during the tetragonal to orthorhombic distortion as the specimen is cooled. In adjacent domains the $[100]$ and $[010]$ directions alternate, thereby allowing relaxation in a given parent grain of the stress resulting from the difference of $\sim 1.6\%$ between the a and b lattice parameters caused by oxygen-site ordering (see, for example, refs 9, 13). Although the c axis is accurately maintained from domain to domain (as shown in Fig. 1*b*), the slight misorientations ($\leq 1^\circ$) of the $[100]$ and $[010]$ directions in adjacent domains are consistent with the model. If this description is correct then the optimum domain boundary spacing is expected to vary both with the parent grain size and with oxygen deficit as the a/b ratio changes. Boundaries with $\{110\}$ habit separating similarly misoriented domains have been observed previously and described as 'twin' boundaries^{17,18}. $\{110\}$ habit boundaries are to be expected because they can be coherent, and in this state the angle between b and a in adjacent grains allows a measurement of b/a . In the material examined the $\{110\}$ boundaries contained defects and the misorientation was less than would be expected for the b/a ratio⁸. The reason for this is not understood but would correlate with the paralleled tendency for the oxide examined to exhibit domain habits other than $\{110\}$. The different habits observed from grain to grain were in a decreasing abundance in the sequence $\{113\}$, $\{100\}$, $\{110\}$. The characteristic domain boundary spacing in our high- T_c material is in the range 10 – 100 nm, although Fig. 2 shows the characteristic arrowhead-shaped nucleation of a small intermediate domain

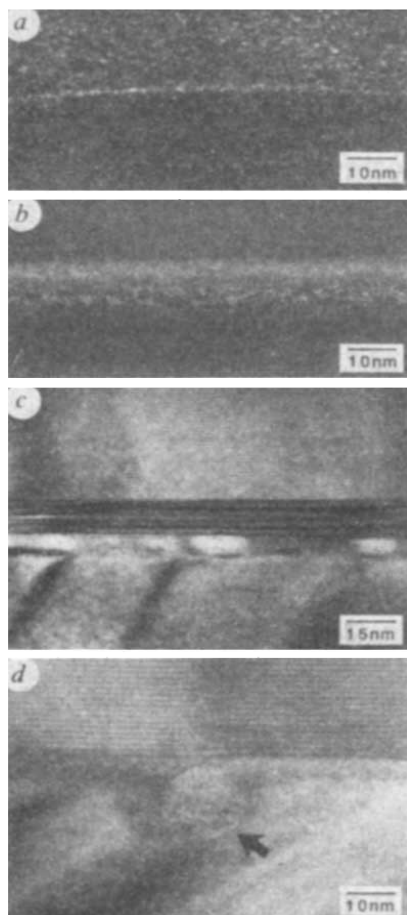


Fig. 3 *a, b*, Dark-field images of amorphous parent-grain-boundary interlayers. *c*, Gross modification of the *c* spacing next to the amorphous boundary layer and *d*, cusping of the domain boundaries, indicative of the enhanced oxygen transporting properties of these boundaries. *c, d* Are bright-field images. In *d*, the arrow points to a second-phase impurity particle.

from a parent grain boundary in a region where the boundary habit is close to {100}.

In the sample studied, up to 40% of the parent grain boundaries were associated with a variety of different structural modifications, including the formation of second phases and bands of amorphous material (Fig. 3), although the remainder exhibited no discernible change in the oxide structure. The amorphous intergrain boundary layers, when observed, had thicknesses ranging from 1 to 5 nm (Fig. 3). Figure 3*c* shows a gross modification of the *c* spacing in one grain and cusping at the domain boundaries in the adjacent grain, indicating modification of oxygen transport¹⁹. Cusping of the domain boundaries is again seen in Fig. 3*d*, which also shows a second-phase impurity particle. At boundaries such as these there proved to be yttrium segregation associated with the presence of segregated silicon impurities, probably picked up during grinding in an agate ball mill. At other boundaries where there was a structural modification but where no silicon was observed there was significant barium segregation. Critical-current and magnetization measurements on wire and bulk samples fabricated from these powders confirmed both the presence of a relatively large local critical current density of $\sim 10^6 \text{ A cm}^{-2}$ at 4.2 K in zero magnetic field⁵ and the dramatic effect of silicon additions in degrading the critical current density⁹. Figure 4 shows a critical-current transition for a reference wire (*a*), and for an identical wire prepared simultaneously but with 0.1 wt % added SiO_2 impurity (*b*). These results are consistent with the general propensity of ceramic materials both to form grain-

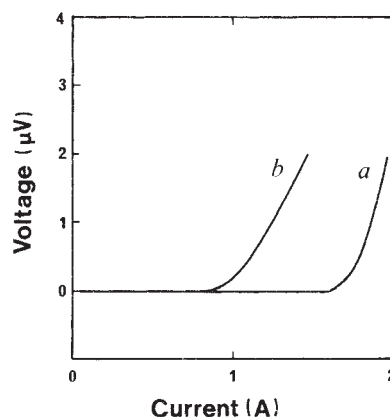


Fig. 4 Critical-current transitions for superconducting wires of cross-section 0.36 mm^2 : *a*, wire made from powder as described in text (critical current density 445 A cm^{-2}); *b*, wire made from powder with 0.1 wt % added SiO_2 (233 A cm^{-2}).

boundary glassy phases and to exhibit strong grain-boundary segregational effects. As a consequence the microstructure and properties of oxide superconductors are expected to depend strongly on the level of trace impurities, as demonstrated here for silicon.

The microstructure described thus has features that can explain the observed critical-current variation in sintered $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. The long-range critical transport current is generally low, and decreases smoothly with increasing temperature over the whole range from 4 K to the critical temperature. In addition, the critical current falls sharply with increasing magnetic field, although at 4 K a small constant critical current has been observed in fields up to 12 T (ref. 20). In contrast, measurement of the magnetization of powdered material of different sizes has clearly indicated that persistent currents in excess of 10^6 A cm^{-2} at 4 K are supported locally in certain regions within the microstructure^{4,5}; similar current densities have been reported for epitaxial thin films⁶.

Three distinct phenomena have to be considered: flux-line pinning at domain boundaries within the grains; flux-line pinning at grain boundaries; and proximity coupling or weak-link behaviour between adjacent grains separated by semiconducting or normal metal boundary layers. It is likely that the domain substructure will pin flux vortices strongly, the pinning possibly varying as a function of the habit. In $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ there are likely to be two contributions to the pinning force. Any basal-plane anisotropy of the upper critical field will lead to pinning; an anisotropy of 1% could lead to the observed local critical currents (ref. 7, p. 177). In addition, the interruption of the linear Cu-O chains by domain boundaries is likely to depress T_c near a boundary, and T_c is also likely to be depressed within any domain containing interrupted chains. Both effects will lead to strong pinning. Pinning at parent grain boundaries due to anisotropy is likely to be very large, if the oxide structure is undisturbed to within less than a superconducting coherence length, ξ , of the physical boundary. (In these materials ξ appears to be of the order of 2 nm (ref. 21).) But layers of semiconducting or normal metallic material greater than ξ in thickness will have a catastrophic effect on current transport from grain to grain. The observed wide variation of critical current with temperature indicates that many of the grain-boundary layers become superconducting as the temperature is lowered. Equally, the strong variation of critical current with magnetic field is an indication of current limitation by tunnelling across extensive semiconducting or normal metal layers.

The domain boundaries described here will act as effective pinning sites in both sintered material and epitaxially grown and constrained single-crystal material, in which they would also be expected to be formed. In the future development of

the material it will be important to control the formation and thickness (to less than ξ) of any amorphous interlayers between the parent grains, as boundaries much thicker than this will catastrophically limit the critical current. It is possible that this could be achieved by the addition of controlled amounts of selected grain-boundary glass-forming segregants. This might allow the retention of the beneficial grain-boundary property of flux pinning without unduly reducing the critical current.

Received 25 June; accepted 18 August 1987.

1. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
2. Wu, M. K. *et al. Phys. Rev. Lett.* **58**, 908–909 (1987).
3. Zhao, Z. X. *et al. Kexue Tongbao*, **10**, 661–663 (1987).
4. Suenaga, M., Ghosh, A., Asano, J., Sabatini, R. L. & Moodenbaugh, A. R. in *Proc. Mater. Res. Soc. Meet.*, Anaheim, April 1987 (in the press).
5. Campbell, A. M. *et al. in Proc. 1st Eur. Workshop, High-T_c Superconductors and Potential Applications*, Genova, July, 1987 (in the press).
6. Laibowitz, R. B., Koch, R. H., Chaudhari, P. & Gambino, R. J. *Phys. Rev. Lett.* **35**, 8821–8824 (1987).
7. Campbell, A. M. & Evetts, J. E. *Critical Currents in Superconductors* (Taylor and Francis, London, 1972).
8. Camps, R. A., Evetts, J. E., Glowacki, B. A., Newcomb, S. B. & Stobbs, W. M. *J. Mater. Res.* (in the press).
9. Glowacki, B. A. & Evetts, J. E. in *Proc. 1st Eur. Workshop, High-T_c Superconductors and Potential Applications*, Genova, July, 1987 (in the press).
10. Capponi, J. J. *et al. Europhys. Lett.* **3**, 1301–1308 (1987).
11. David, W. I. F. *et al. Nature* **327**, 310–312 (1987).
12. Monod, P. *et al. J. Phys.*, Paris (in the press).
13. Jorda, J. L., Graf, T., Bezing, A., Junod, A. & Müller, J. in *Proc. 1st Eur. Workshop, High-T_c Superconductors and Potential Applications*, Genova, July, 1987 (in the press).
14. Ourmazd, A. *et al. Nature* **327**, 308–310 (1987).
15. Hewat, E. A., Dupuy, M., Bourret, A., Capponi, J. J. & Marezio, M. *Nature* **327**, 400–402 (1987).
16. Hyde, B. G. *et al. Nature* **327**, 402–403 (1987).
17. Syono, Y. *et al. J. appl. Phys.* **26**, L498–L501 (1987).
18. Larbalestier, D. C. *et al. J. appl. Phys.* (submitted).
19. Newcomb, S. B., Metcalf, E. & Stobbs, W. M. *Phil. Trans. R. Soc. A* **319**, 191–218 (1986).
20. Yoshino, H. *et al. Pap. presented at Eur. Mater. Res. Meet.*, Strasbourg, June 1987.
21. Cava, R. J. *et al. Phys. Rev. Lett.* **58**, 1676–1679 (1987).

Solution of phase problem for crystallography at a wavelength of 3.5 Å

W. Shen & R. Colella

Purdue University, Physics Department, West Lafayette, Indiana 47907, USA

It has been established¹ that one viable method for determining phases in a diffraction experiment is to make use of multiple beam-diffraction, that is, a situation in which two or more reflections are excited simultaneously in a crystal. While the theory had been fully developed and tested in some experiments on a germanium crystal¹, progress was halted by the recognition that such theory could only be applied to perfect crystals such as germanium and silicon, with no mosaic structure. Here we report that considerable progress has been made, to the point of being able now to determine phases of structure factors in the most general situations, with non-centrosymmetric, small mosaic crystals of arbitrary shape. An example will be described here in which the phase of the (202) reflection of an organic molecular crystal (benzil) has been determined.

Important progress was made in 1981², when it was shown that certain anomalies observed on the wings of azimuthal plots contain phase information, and extend to wide angular ranges, of the order of degrees. (By azimuthal plot we mean a plot in which each point is an integrated intensity of the (*hkl*) reflection, integrated with respect to θ , angle of incidence on the (*hkl*) lattice planes, as a function of ϕ , the azimuthal angle of the crystal with respect to the scattering vector.) It was found, in ref. 2, that a new mechanism of diffraction is operative in situations in which the main reflection is weak and fully excited, whereas the simultaneous reflection is strong but weakly excited. Such mechanism has been called 'virtual Bragg scattering', (VBS), in analogy with the concept of virtual transitions in atomic and nuclear physics (A. W. Overhauser, personal communication).

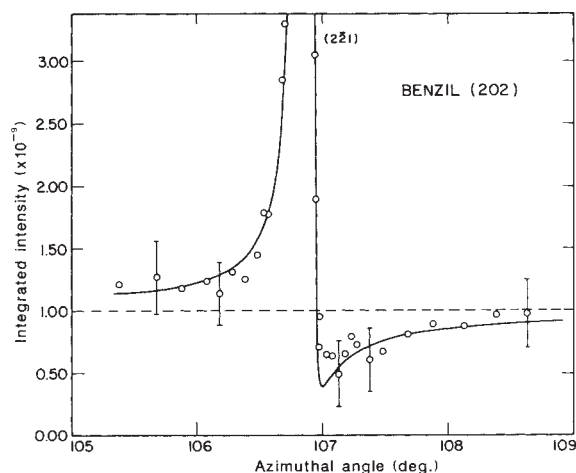


Fig. 1 Azimuthal plot of the (202) reflection, around the excitation point of (221). The dotted line corresponds to the two beam value. The counting rate of the (202) peak value, in the two beam region, was ~ 1 count s^{-1} . The total number of counts accumulated for each data point was around 1,400 of which 1,000 was background. The solid line is given by a six-beam calculation, using perturbation theory⁷, and a smearing function is introduced to take into account a 0.03° mosaic spread. More details are given in ref. 6.

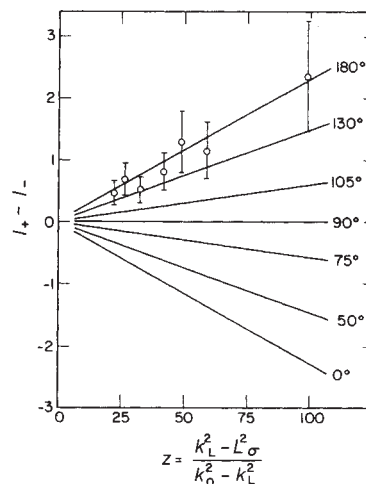


Fig. 2 Plot used to determine $\phi_{221} + \phi_{021} - \phi_{202}$. The ordinate of each point is the difference in integrated intensity, normalized to the two-beam value, between two experimental points in Fig. 1, symmetrically located with respect to the (221) peak. (The variable z , plotted on the horizontal axis, is a geometrical factor which increases as the excitation point of the (221) is approached, and is explained in ref. 6.)

It was immediately recognized that, since the VBS situation is one in which the scattering is kinematic (no multiple scattering), it could be used to determine phases in mosaic crystals, by using dynamical theory, the only one that preserves phase information.

As a first step, to test the soundness of our ideas and methods, we confirmed by VBS the sign change³ of the 'forbidden' (442) reflection in silicon as a function of temperature⁴.

The second crucial step, was the phase determinations of some 'forbidden' reflections in V_3Si , a mosaic crystal⁵, the first time in which multiple-beam diffraction was used to determine phases *a priori* unknown.

At this point we turned to an organic, molecular, non-centrosymmetrical crystal (benzil: $C_{14}H_{10}O_2$) to check the feasibility of our method on a typical representative of a class of crystals for which phase problems are normally encountered. Initially we were not able to see any VBS effect, even when a highly collimated beam from a synchrotron source was used. A calculation of the expected azimuthal profile, for 8 keV X-rays,

showed that the asymmetry effects for a perfect benzil crystal would be confined to an angular range, on the azimuthal scale, of less than 2 arc min, which is about the mosaic spread of our crystal. It became clear that, irrespective of the instrumental resolution, no VBS effects could be observed in a crystal like benzil. The different behaviour of benzil, compared to V_3Si or Si , is probably due to the fact that only light atoms are present in its unit cell.

Further calculations showed that by increasing the X-ray wavelength VBS effects could be made visible over wider angular ranges. For example, our calculations showed that at a wavelength of 3.5 Å pronounced asymmetry effects were visible in the azimuthal plots of the (202) reflection in benzil over angular ranges of 1–2°.

An experiment was done at the National Synchrotron Light Source (NSLS: Brookhaven National Laboratory) in which a large, thick slab of benzil, cut parallel to the (202) planes, completely intercepting the beam, exhibited the desired effect⁶. Using the perturbation theory⁷ of multiple-beam diffraction developed by one of us (Q.S.), the experimental points could be plotted along a straight line whose slope is directly related to the phase of the reflection involved.

One problem that, we felt, might have prevented widespread application of our method, was the use of boundary conditions in the general theory¹ developed in 1974. It was felt, however, that, given the kinematic nature of VBS, the boundary conditions should not play any role, and that a crystal of any shape should give the same results, irrespective of shape, apart from absorption effects. (The absorption correction is not negligible at 3.5 Å. But the angular ranges over which the crystal is rotated in an azimuthal plot are quite small (3–4° at most). This means that the absorption correction is the same for all points and can hence be neglected.) Our conjecture was reinforced by the

rigorous perturbation treatment of ref. 7, in which no use is made of boundary conditions.

A recent experiment, performed on the X-ray 8-pole wiggler line VII-2 at the Stanford Synchrotron Radiation Laboratory, has fully confirmed our expectations. A small quasispherical crystal of benzil (0.3-mm diameter) was set for the (202–221) reflection. The data are shown in Fig. 1, which is essentially identical to Fig. 4 of ref. 6, in which a flat crystal slab was used. Figure 2 shows how the data can be plotted, using the theory of ref. 7, to get the sum of the phases of (221) and (021) with respect to (202), which turns out to be 180°, as expected. The scatter of the data points is due to poor counting statistics, due to the low intensity of the 3.5 keV photons. With the new generation of synchrotron sources and multipole insertion devices and by further reducing the air paths, the counting rates can be easily increased by at least two orders of magnitude. (At 3.5 keV an air path of 30 cm transmits only 1%. In our last experiment at SSRL we had a total of ~21 cm of air paths in the hutch, half of which can certainly be eliminated by constructing special collimators.)

We have therefore confirmed that our method, based on the notion of virtual Bragg scattering, is universally applicable and can be used to solve unknown structures, with no restrictions.

This work was supported by the National Science Foundation and by the Department of Energy.

Received 17 March; accepted 16 June 1987.

1. Colella, R. *Acta crystallogr.* **A30**, 413–423 (1974).
2. Chapman, L. D., Yoder, D. R. & Colella, R. *Phys. Rev. Lett.* **46**, 1578–1581 (1981).
3. Trucano, P. & Batterman, B. W. *Phys. Rev.* **B6**, 3659–3666 (1972).
4. Tischler, J. Z., Shen, Q. & Colella, R. *Acta crystallogr.* **A41**, 451–453 (1985).
5. Schmidt, M. C. & Colella, R. *Phys. Rev. Lett.* **55**, 715–717 (1985).
6. Shen, Q. & Colella, R. *Acta crystallogr.* (submitted).
7. Shen, Q. *Acta crystallogr.* **A42**, 525–533 (1986).

Palaeomagnetic data suggest link between the Archaean–Proterozoic boundary and inner-core nucleation

C. J. Hale

Department of Geology, McMaster University, Hamilton, Ontario, Canada L8S 4MT

It is generally believed that the Earth's magnetic field arises from dynamo action in the convecting outer core^{1–3}. Thus the geomagnetic field, and especially its strength, can provide information about core processes. Recent determinations of the geomagnetic palaeointensity using Archaean and early Proterozoic samples from the Kaapvaal and Superior cratons are consistent with a sharp increase in the magnitude of the geomagnetic field between 2.7 and 2.1 Gyr. Analysis of published late Archaean and early Proterozoic apparent polar wander paths shows that an abrupt change in the rate and style of continental motion occurred at about 2.5–2.6 Gyr ago. If this change is also interpreted to be related to the growth of an inner core, then, an intriguing conjecture arises that the Archaean–Proterozoic transition is related to inner core nucleation.

Most current theories of core formation imply that the differentiation of a nearly homogenous, cold-accreted Earth into a radially-stratified body with a silicate mantle and iron-rich core took place within a few hundred million years of accretion. Many different energy sources have been considered to provide the necessary heat for melting and separation of the iron fraction. These range from the gravitational energy of the accreting planetoids^{4,5} to short-lived radioisotopes such as ²⁶Al (ref. 6) and superheavy elements⁷.

Once a temperature was reached which permitted the differentiation process to begin, it probably proceeded rapidly

since additional energy became available from the gravitational energy of the sinking iron. This energy, amounting to 500 or 600 calories per gram (refs 8, 9) would be sufficient to raise the whole earth temperature by 1,500 or 2,000 K unless a substantial amount of heat was lost by radiation at the Earth's surface. The question of the timing of core formation reduces to a question of the balance between the rates of heat production and transport out of the early Earth.

This transport must have involved convection in the core. As there was no shortage of energy to drive a dynamo in the early Earth, it is reasonable to suppose that the geomagnetic field has been present almost since terrestrial accretion.

It is more difficult to explain why the geomagnetic dynamo is still operative. A problem arises since no heat is available from sources such as gravitational energy and short lived isotopes to maintain the geomagnetic dynamo beyond about 1.0 Gyr ago³. One plausible explanation lies in the continuous growth of the inner core which provides enough gravitational energy to keep the geomagnetic dynamo "stirred"^{2,3,10,11}.

Palaeomagnetic data can provide a firm constraint on the timing of core formation since the existence of a geomagnetic field implies that the Earth is differentiated and that its core is convecting. As the geodynamo is a consequence of convective motion in the core, the magnitude of the field is related to the core's rate of convection. Determinations of the strength of the ancient geomagnetic field can provide experimental evidence against which to test some core-dynamo models. For example, the thermal evolution models of Stevenson *et al.*¹², which indicate that the inner core nucleated as recently as 1.5–2.5 Gyr ago, also require an abrupt change in the palaeointensity at about that time.

The geomagnetic field at 3.47 Gyr was weaker than at present^{13,14} with an equatorial palaeointensity of <0.1 Oe, compared to the present intensity of 0.3 Oe. Thellier and Van Zijl experiments performed on samples from the type locality of the Komati

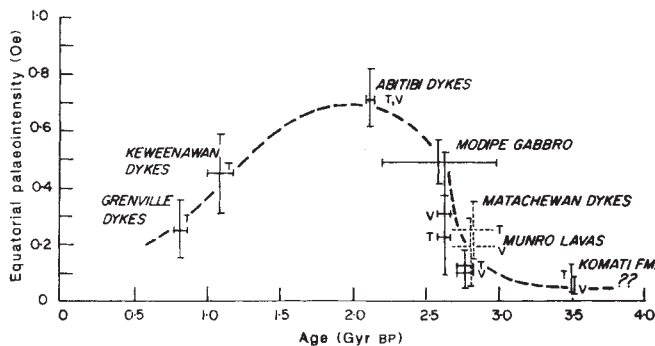


Fig. 1 Palaeointensity values plotted against age. The results from the Komati Formation, Pykes Hill, Matachewan Dykes and Abitibi Dykes come from recent work^{13,14}. The dotted cross indicates the palaeointensity obtained from the Pyke's Hill komatiite samples, calculated after structural correction. The Pyke's Hill data are indistinguishable from the other Munro Lavas if the *in situ* inclination is used to calculate the equatorial equivalent palaeointensity. The Pyke's Hill result has been plotted tentatively near 2.7 Gyr but this age is uncertain. Additional data have been plotted, representing studies¹⁵⁻¹⁷ involving multiple sampling of slowly cooled units.

Formation in the Barberton Greenstone Belt produced similar results. In that study, comparable palaeointensities were recovered from both peridotitic and basaltic komatiites even though their natural remanent magnetism intensities differed by several orders of magnitude.

Archaean samples from the Canadian Shield also record palaeointensities which are low by present-day standards¹⁴. Pyke's Hill komatiite samples give equatorial equivalent palaeointensities of 0.13 and 0.10 Oe (*in situ*) or 0.26 Oe and 0.19 Oe (after structural correction) in Thellier and Van Zijl experiments, respectively. Archaean basalt and gabbro samples yield palaeointensities of 0.12 Oe (Thellier) and 0.10 Oe (Van Zijl) in these experiments. Again, palaeointensities obtained from strongly magnetic peridotitic komatiites and weakly magnetized basalt samples agree reasonably well as do the results of the two different methods of determination.

Palaeointensities have also been obtained from the Matachewan diabase dykes and their baked contacts (0.22–0.3 Oe), more like the present-day equatorial geomagnetic field intensity. These values agree reasonably well with the determinations from individual Proterozoic samples which led Schwarz and Symons¹⁵ to the conclusion that there has been no long-term variation in the geomagnetic field intensity since 2.5 Gyr ago.

In contrast to that view, however, 2.15-Gyr Abitibi dyke samples have yielded palaeointensities which are over twice the present geomagnetic field intensity at about 0.7 Oe (ref. 14). While this high value was determined less precisely than the older data, it is not unreasonable. McElhinny and Evans¹⁶ also obtained a palaeointensity 1.5 times the present value from the

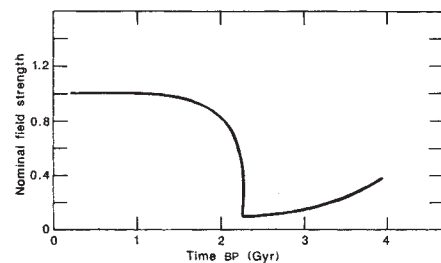


Fig. 2 Nominal geomagnetic field strength throughout geological time as calculated by Stevenson *et al.*¹². Although this calculation provides only a broad estimate of the dipole field strength, there exists a remarkable similarity between the predicted field behaviour and the experimental data.

late Archaean or early Proterozoic Modipe gabbro. Pesonen and Halls¹⁷ determined high palaeointensities (1.5 times the present intensity) from 1.0 to 1.2 Gyr Keweenaw dykes and lavas.

To date, these are the only Precambrian palaeointensity studies which involve multiple sampling of slowly-cooled intrusions in which short-term variations in the geomagnetic intensity are averaged out. The range of averages obtained may thus reflect real differences in the long-term strength of the geomagnetic dynamo.

When these few data are plotted against their ages as in Fig. 1, a striking pattern emerges. Sometime before the intrusion of the Matachewan dykes, the field began a dramatic growth, reaching a maximum intensity during the mid-Proterozoic. Inference of this pattern relies heavily on the high palaeointensities obtained from the Abitibi dyke samples. Since long gaps at 1.5–2.0 Gyr and 2.2–2.5 Gyr leave the pattern open to doubt, the direction for future work is obvious. Recognizing that the exercise is perhaps premature, one may still examine the implications of such a pattern.

Figure 2 is redrawn after Stevenson *et al.*¹² who equated the energy available for dynamo generation to the Ohmic dissipation of energy in the core in order to predict a nominal magnetic field intensity for the Earth. Their model predicts a low geomagnetic field intensity prior to the nucleation of the earth's inner core sometime between 1.5 and 2.5 Gyr ago. McElhinny and Evans' palaeointensity value from the Modipe gabbro¹⁶ was cited as evidence against the model but then these data were ignored in a sweeping condemnation of palaeointensity studies as uncertain.

The experimental results highlighted above lend support to the theory of Stevenson *et al.*¹², except as to the timing of inner core nucleation. The apparent conflict with McElhinny and Evans' Modipe result probably reflects the uncertainty in both the age of the gabbro and the timing of inner core nucleation rather than a fundamental problem with the model. The experimental data are quite consistent with theory if inner-core nucleation is interpreted to have begun at ~2.7 Gyr rather than

Table 1 Palaeomagnetic directions used to calculate Archaean poles for North America and the implied palaeolatitudes

Unit	Age (Gyr)	Dec./Inc.*	α_{95}	Palaeolatitude	dP	Reference	
			degrees				
Pykes Hill lavas	>2.7	178/−12	6	6	3	PH	25
Archean gabbros	<2.7	189/−23	17	12	10	AB	25
Archean basalts	>2.7	28/+25	11	13	6	AB	26
Skead pyroclastics	>2.7	218/+19	10	10	4	SK	27
Ghost Range complex	2.710–2.703	280/+2	12	1	6	GR	28
Blake River volcanics	2.703	190/−29	25	15	15	BR	29
Adams Mine A component	2.7	270/+9	—	4		AM	30
Kamiskotia complex	>2.6	289/+23	5	12	3	KK	25
Dundonald sill	>2.6	269+32	18	17	11	DS	25
Matachewan dykes	2.63	207/−16	7	8	4	MD	25
		21/+13	6	7	3	MD2	26
Wabigoon gabbro	2.6	246/+12	11	6	6	WB	30

* Declination and inclination.

2.5 Gyr. Given the approximate nature of the model of Stevenson *et al.*, this change is minor.

New palaeointensity data thus support the idea that a rapid increase in the intensity of the geomagnetic field occurred at about the time of the Archaean-Proterozoic boundary. If the model of Stevenson *et al.* is essentially correct, then this increase may be attributed to the nucleation of the Earth's inner core. It is particularly intriguing to explore the idea that the profound transition marking the close of the Archaean is itself related to inner-core nucleation. As cratonic crust was stabilized on a global scale at that time, it is reasonable to suppose that an event of global proportions took place. It is equally reasonable to expect that such a profound change in the structure of the earth would be accompanied by a change in apparent polar wander paths (APWPs).

Convection in the core provides heat to the bottom of the mantle and this heat must be transported by mantle convection. Thus, after an interval required to propagate a thermal disturbance to the surface, the rate of continental drift (which marks the upper limb of mantle convection) can be expected to reflect changes in core convection. The extensive plutonism and crust-building which took place between ~3.0 and 2.8 Gyr (ref. 18) may be related to enhanced mantle convection, cooling the core and permitting the nucleation of the inner core. Stevenson *et al.*¹² point out that the thermal history of a terrestrial planetary core is governed by the ability of the mantle to cool the core. They conclude that the history of the magnetic field is tied to the thermal history of the core and is therefore intimately related to mantle convection.

This idea is not a new one. For example, Vogt¹⁹ tried to correlate magnetic field reversal rates with changes in the rates of sea-floor spreading deduced from the Atlantic and Pacific sea-floor anomaly patterns. More recently, McFadden and Merrill²⁰ have argued that Phanerozoic reversal rates reflect thermal changes at the core-mantle boundary with time constants of ~10⁸ yr. At least part of the problem with these correlations arises from the difficulty in estimating the time lag between core processes and their surface manifestation. The apparent polar wander signature of a large, unique event such as the nucleation of the inner core, on the other hand, ought to be accompanied by an unmistakable change in the rate of apparent polar wander, and this may permit an estimate of the time lag.

Figure 3a, redrawn after Hale and Dunlop²¹, shows that Archaean data from southern Africa cluster in a comparatively small part of the globe. There is thus little evidence of apparent polar wander. While the record is far from complete, the proximity of the Komati and Modipe poles suggests that either southern

Africa remained in a high-latitude position for 900 Myr or it returned there. Figure 3b is redrawn after McElhinny and Senenayake²² with the omission of their Duffer Formation pole which is probably a Proterozoic overprint. The tight grouping of Archaean and early Proterozoic poles is evidence of a persistent high-latitude position for ancestral Australia in the late Archaean.

Figure 4 reproduces Dunlop's²³ apparent polar wander path for North America in the late Archaean. At first glance, this path seems quite unlike the African and Australian data since it consists of a few pole positions spread out along a long swathe. Table 1 lists the palaeolatitudes implied by the characteristic directions for poles which predate the intrusion of the

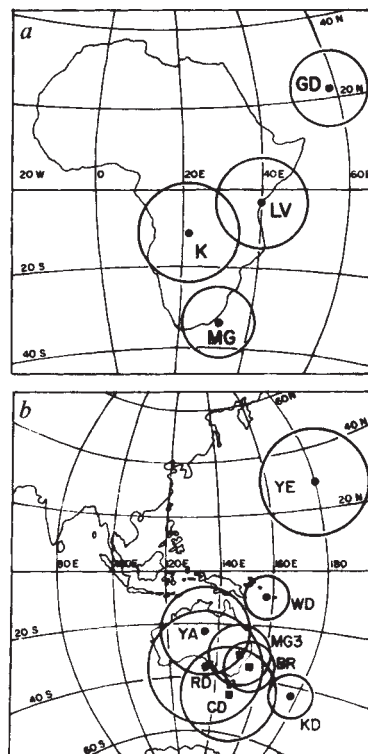


Fig. 3 Archaean and early Proterozoic pole positions from Africa and Australia¹³ show a remarkable tendency to group in the vicinity of the present-day continent. This is evidence of a persistent high-latitude position for both ancestral continents. The palaeolatitudes which correspond to these poles are shown in Table 2.

Table 2 Palaeomagnetic directions used in calculating Australian and African Archaean poles and the implied palaeolatitudes

Unit	Age (Gyr)	Dec./Inc.	α_{95}	Palaeolatitude	dP	Reference
Australian poles						
Yilgarn Craton						
Yilgarn	2.5	61/+81	10	72	18	YD 32
Dykes						
Ravensthorpe Dykes	2.5	114/+83	13	76	25	RD 32
Widgemooltha Dykes	2.42	242/-67	6	50	8	WD 33
Koolyanobbing Dowd's Hill	2.75-2.2	303/-61	7	41	9	KD 34
Pilbara Craton						
Mt Goldsworthy ores	3.0-2.0	115/+74	9	59	13	MG3 34
Black Range Dyke	2.33	115/+72	6	57	9	BR 35
Cajaput Dyke	2.4	145/+71	13	55	22	CD 35
African poles						
Komati Fm	3.5	326/+81	8	72	15	21
Modipe gabbro	2.6	155/+85	5	80	10	36
Lower Ventersdorp lavas	2.5	35/+75	7	62	11	37*
Great Dyke	2.5	220/-58	8	39	12	38

dP is uncertainty in palaeolatitude corresponding to the confidence interval (α_{95}) for each pole in Tables 1 and 2.

* Declination and inclination calculated from pole position using latitude and longitude of Ventersdorp.

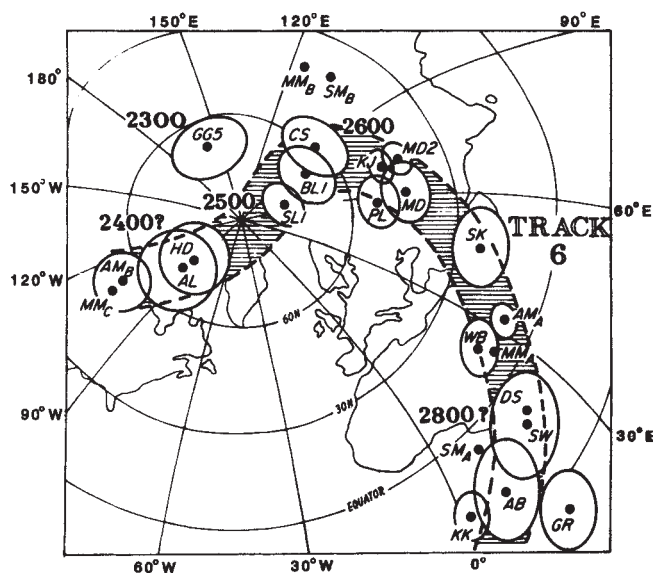


Fig. 4 Late Archaean and early Proterozoic APWP for Laurentia as drawn by Dunlop²³. Poles between ~2.8 and 2.6 Gyr lie along a long track but all of these remanences were acquired at low latitudes (see Table 1). Sometime between 2.5 and 2.6 Gyr, the APWP turns, becoming radial with respect to North America. This is evidence of poleward continental displacement and is the first such evidence.

Matachewan dykes at 2.63 Gyr. Only the Dundonald Sill gave a palaeolatitude higher than 15°. The entire range of palaeolatitudes is only 16°, about the uncertainty in a typical Precambrian APWP width. Since palaeolongitude changes are not shown by palaeomagnetic data, one cannot rule out the possibility of east-west continental motion for North America at the close of the Archaean but poleward motion is excluded by the data. The entire sweep of the 2.8–2.6 APWP may be explained by a simple rotation of ancestral North America about a local vertical axis.

Such rotations of continental fragments are well documented for Phanerozoic poles from western North America²² and are known to be common feature of plate boundaries. On the other hand, the parallelism of modern seafloor magnetic anomalies and the maintenance of the relative orientations of Africa and South America argue that at present continents do not tend to rotate about interior Euler poles during quiescent spreading. Rotation of the North American continent in low latitudes at the close of the Archaean suggests that this may have been an important boundary of the prevailing mantle convection regime. The pre-2.6-Gyr Archaean APWP for North America thus can be explained without continental drift.

The Australian and African data are particularly informative because of their high-latitude clustering. In these cases, any continental motion must necessarily involve a change in palaeolatitude. To date, available palaeomagnetic data do not require large scale motion of either of these continents at the close of the Archaean. The Archaean situation is thus quite unlike the pattern of Proterozoic APWPs which are characterized by long tracks indicating palaeolatitude change, separated by hairpin bends and episodes of rotation²⁴.

If the nucleation of the Earth's inner core is tied to enhanced mantle convection, this enhancement may be shown by a transition from fixed continents in the late Archaean to mobile continents in the Proterozoic. A possible timing for this transition is suggested by the abrupt change in the direction of the Archaean-Proterozoic APWP for North America which occurs between 2,500 and 2,600 Myr.

If one adopts this date for an Archaean-Proterozoic transition, there is a time lag of ~150 Myr between the palaeointensity increase at 2.7 Gyr which heralds nucleation of the inner core and the acceleration of continents at ~2.55 Gyr.

The high-latitude clustering of the Australian and African pole positions is intriguing in itself. Table 2 lists the palaeolatitudes which correspond to these poles. All of the poles older than 2.5 Gyr indicate palaeolatitudes higher than 50°. Since less than 24% of the Earth's surface occupies latitudes higher than 50°, this abundance of high latitude poles invites a search for some mechanism to explain the high-latitude clustering.

Close grouping of the North American data in low latitudes suggests that this continental nucleus remained near an equatorial position at the close of the Archaean. Both polar and equatorial positions were apparently favoured for long periods.

The Archaean-Proterozoic boundary marks a singularly important event in the history of the Earth. Apparent stability of continent positions in the Archaean gave way to the relative mobility of continental nuclei shown by the Proterozoic APWP tracks.

A rapid increase in the geomagnetic field strength also occurred during the late Archaean and early Proterozoic, consistent with the predictions of the thermal model of Stevenson *et al.*, of a late nucleation of the terrestrial inner core. These separate observations are tied together by the conjecture that profound geological changes at the Archaean-Proterozoic boundary may be related to the nucleation of the inner core. This idea can be tested in three ways:

First, more truly Archaean palaeomagnetic pole positions are needed. According to the model presented above, these ought to imply palaeolatitudes which are either equatorial or greater than about 50°. They ought to be consistent with the positions defined by the few presently available Archaean data, requiring little continental drift. Second, more Archaean palaeointensities are needed. Remanences older than the 3.5-Gyr Komati Formation characteristic remanence should indicate higher palaeointensities. Remanences younger than the Komati but older than ~2.8 Gyr should indicate very low palaeointensities as this is presumably the time at which core convection became sufficiently sluggish that differentiation could proceed. Finally, more Proterozoic palaeointensities are required. Long-term measures of the geomagnetic palaeointensity can be compared with the pattern presented in Fig. 1. Data in the critical period between 2.7 and 2.1 Gyr are needed to establish or refute the hypothesized increase in palaeointensity and its link with inner core nucleation.

This work has been supported under an operating grant from the Natural Sciences and Engineering Research Council of Canada. Many of the measurements were carried out at the University of Toronto Paleomagnetism Laboratory, aided by an NSERC operating grant to Professor D. J. Dunlop whose support and encouragement is also acknowledged.

Received 18 February; accepted 10 July 1987.

- Jacobs, J. A. in *Physics of the Earth's Interior* (Soc. Italiana di Fisica, Bologna, 1980).
- Gubbins, D. *Contemp. Phys.* **25**, 269–290 (1984).
- Stevenson, D. J. *Rep. Prog. Phys.* **46**, 555–620 (1983).
- Safranov, V. S. *Nauka, Moscow* (1969); transl. NASA TTF-667 (1972).
- Kaula, W. *Geol. Soc. Can. Spec. Pap.* No. 20 (1980).
- MacDonald, G. J. F. *J. geophys. Res.* **67**, 167 (1959).
- Runcorn, S. K. *Nature* **270**, 676 (1978).
- Tozer, D. C. *Geophys. J.* **9**, 95 (1965).
- Flaser, F. M. & Birch, F. *J. geophys. Res.* **78**, 6101 (1973).
- Braginsky, S. I. *Zh. Eksper. i Teor. Fiz.* **47**, p. 2178 *Sov. Phys. JETP English Translation* **20** (1965), 1462 (1964).
- Gubbins, D. *J. geophys.* **43**, 453 (1977).
- Stevenson, D. J., Spohn, T. & Schubert, G. *Icarus* **54**, 466–489 (1983).
- Hale, C. J. thesis, Univ. Toronto (1985).
- Hale, C. J. *Earth planet. Sci.* (in the press).
- Schwarz, E. J. & Symons, D. T. A. *Phys. Earth planet. Int.* **2**, 11–18 (1969).
- McElhinny, M. & Evans, M. E. *Phys. Earth planet. Int.* **1**, 485–497 (1968).
- Pesonen, L. & Halls, H. C. *Geophys. J. R. Soc.* **73**, 241–270 (1983).
- Nelson, B. K. & DePaolo, D. J. *Geol. Soc. Am.* **96**, 746–754 (1985).
- Vogt, P. R. *Earth planet. Sci. Lett.* **25**, 313–321 (1975).
- McFadden, P. L. & Merrill, R. T. *J. geophys. Res.* **89**, 3354–3362 (1984).
- Hale, C. J. & Dunlop, D. J. *Geophys. Res. Lett.* **11**, 97–100 (1984).
- McElhinny, M. W. & Senenayake, W. E. *J. geophys. Res.* **85**, 3523–3528 (1980).
- Dunlop, D. J. *Can. J. Earth Sci.* **21**, 869–878 (1984).
- Beck, M. E. Jr. *Am. J. Sci.* **276**, 694–712 (1976).
- Irving, E. & Naldrett, A. J. *Geology* **85**, 157–176 (1977).
- Schatts, L. D. & Dunlop, D. J. *Nature* **291**, 642–645 (1979).
- Ridley, R. H. & Foster, J. H. *Geol. Surv. Can. Pap.* No. 77–10 (1981).

28. Geissman, J. W. M., Strangway, D. W., Tasillo-Hirt, A. & Jensen, L. S. *Can. J. Earth Sci.* **19**, 2085–2099 (1982).
29. Tasillo-Hirt, A. M., Geissman, J. W., Strangway, D. W. & Jensen, L. S. *Can. J. Earth Sci.* **19**, 2100–2113 (1982).
30. Symons, D. T. A., Quick, A. W. & Stupavsky, M. *Ontario Geol. Survey Misc. Pap. No. 98* 293–307 (1981).
31. Dunlop, D. J. *Can. J. Earth Sci.* **20**, 1805–1817 (1983).
32. Giddings, J. W. *Tectonophysics* **30**, 91–108 (1976).
33. Evans, M. E. *J. geophys. Res.* **7**, 3261–3270 (1968).
34. Porath, H. & Chamalaun, F. H. *Geophys. J.* **15**, 253–264 (1968).
35. Embleton, B. J. *Precamb. Res.* **6**, 275–291 (1978).
36. Evans, M. E. & McElhinny, M. J. *geophys. Res.* **71**, 6053–6063 (1966).
37. Henthorn, D. I. thesis, Univ. Leeds (1973).
38. Jones, D. L., Robertson, I. D. M. & McFadden, P. L. *Trans. Geol. Soc. S. Afr.* **78**, 57–65 (1976).

Increased age estimate for the Lower Palaeolithic hominid site at Olorgesailie, Kenya

Bethany A. Bye*, Francis H. Brown*, Thure E. Cerling* & Ian McDougall†

*Department of Geology and Geophysics, University of Utah, Salt Lake City, Utah 84112, USA

†Research School of Earth Sciences, Australian National University, Canberra 2601, ACT, Australia

The Acheulean is the major stone-tool tradition of Africa, and of much of Asia and Europe, which lasted in Africa from ~1.5 to 0.15 Myr BP. It is associated with *Homo erectus* at some sites, and is characterized especially by hand-axes. Cultural development within the Acheulean has been very hard to assess because so few sites can be dated. Here we report potassium–argon dates and geochemical correlation of tephra layers at Olorgesailie—a key Acheulean site in the Kenya rift valley. The results suggest that the upper part of the Olorgesailie Formation lies between 0.6 and 0.74 Myr BP, and that the maximum age is ~0.9 Myr. Most of the artefact-bearing levels at this site are thus >0.70 Myr old, significantly older than formerly thought.

The Olorgesailie Formation outcrops directly north of Mt Olorgesailie (Oloolkisailie) on the Legemunge (Olekemonge) Plain, in isolated patches to the south-west on the Esonoria Plain, and along the Olkeju Ngiro and its tributaries (Fig. 1). Remarkable accumulations of thousands of stone tools within the formation have been known for some time, and were most recently investigated by Isaac¹, in whose monograph faunal lists, palaeoenvironmental reconstructions, and a discussion of the geology of the area are found.

The volcanic and tectonic history of the Kenyan Rift near Olorgesailie is summarized by Baker and Mitchell², and Baker *et al.*³. The Olorgesailie Formation overlies volcanic rocks of the area, including the nephelinites of Mt Olorgesailie (2.7–2.2 Myr), the Limuru trachyte, pantellerite flood lavas (2.1–1.9 Myr), the Ol Keju Nyiro (also Ol Keju Neru) basalts (2.2–

1.7 Myr), the Ol Tepesi Basalts (1.65–1.4 Myr), and the Plateau and Magadi flood trachyte and pantellerite lavas (1.3–0.9 Myr). Younger alkali trachyte ash cones and scoriaceous lava cones petrologically related to the Plateau Trachytes also occur in the region. The largest example of these is the Ol Doinyo Nyokie (Ol Doinyo Nyegi) ignimbrite complex (Fig. 1), which includes both fissure and central vent derived ash-flow tuffs and unwelded ash layers^{4,5}. K/Ar dates from 0.7 to 0.6 Myr have been obtained for Ol Doinyo Nyokie⁶.

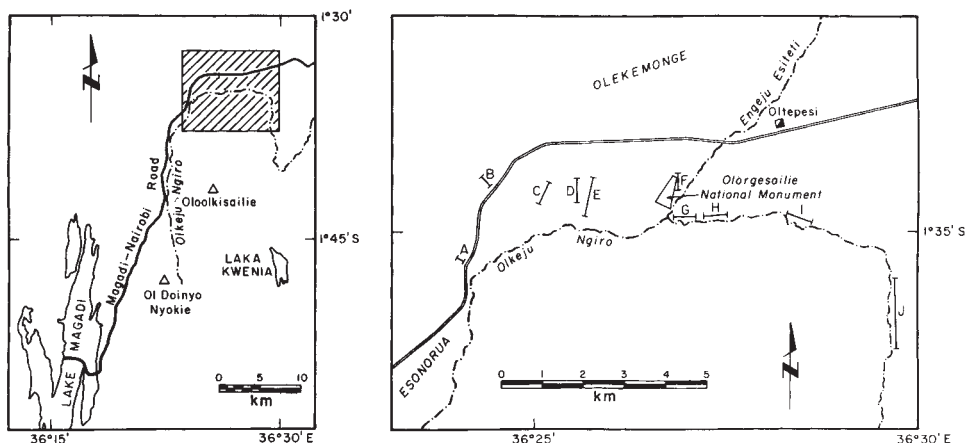
The Olorgesailie Formation is ~60 m thick, and consists mainly of diatomaceous siltstones, with interbedded volcanic ash layers, volcanic sandstones, and pumice rich silts, sands and gravels. Most of the volcanic material was transported to, or reworked within, the sedimentary basin by fluvial processes. Occasional thin airfall tuffs are also present. The formation has been divided into fourteen numbered members^{1,8,9}. Owen and Renault¹⁰ made detailed analyses of the fossil diatom flora, and they also describe the lithofacies and mineralogy of the sediments.

In an effort to refine the stratigraphy, and provide correlations between widely separated sections, tuffaceous horizons and tuffs were sampled at a number of localities. At each locality a stratigraphic section was measured and described to allow lateral correlation with other sections. Volcanic glass separates from tuffs and pumice were produced by standard techniques, and these samples along with whole rock obsidian samples were analysed by X-ray fluorescence spectrometry using methods described by Brown and Cerling¹¹. A detailed discussion of these correlations is given by Bye¹², and the results are summarized in Fig. 2. The correlations provide strong independent support for the stratigraphy worked out by Isaac¹ and Shackleton⁸, and extends that stratigraphy farther to the east.

Representative analyses are given in Table 1, and a bivariate plot of Y against Nb of all tuff samples is shown in Fig. 3 where it is seen that the chemical compositions fall into two clearly separate fields that correspond to different stratigraphic groups. Tuffs from members 1 to 4 form one group, and those of members 8–14 form the second. Because of the clear geochemical separation of the two groups, we suggest that the tuffs represent products of two different volcanic sources. Linear relationships have often been observed between trace element pairs from single volcanic sources^{13–15}. Often elements such as Nb, Y and Zr increase in concentration as a source magma evolves, but reversals of this trend have also been reported. Only one of the samples (B-132) from members 8–14 plots in the field defined by the samples from members 1–4. This may result from a late eruption of a source volcano most active slightly earlier in time, or from reworking of a previously deposited tuff.

No obvious source has been found for the earlier group of tuffs, but the chemical trend defined by the later group of air-fall and reworked tuffs includes the composition of obsidians from Ol Doinyo Nyokie, a faulted volcano situated ~30 km south of the Olorgesailie beds. One tuff from Olorgesailie has a

Fig. 1 Geographical localities of sites and stratigraphic sections described in the text. Place names are from Kenya Survey maps, and do not always correspond to names used in earlier literature.



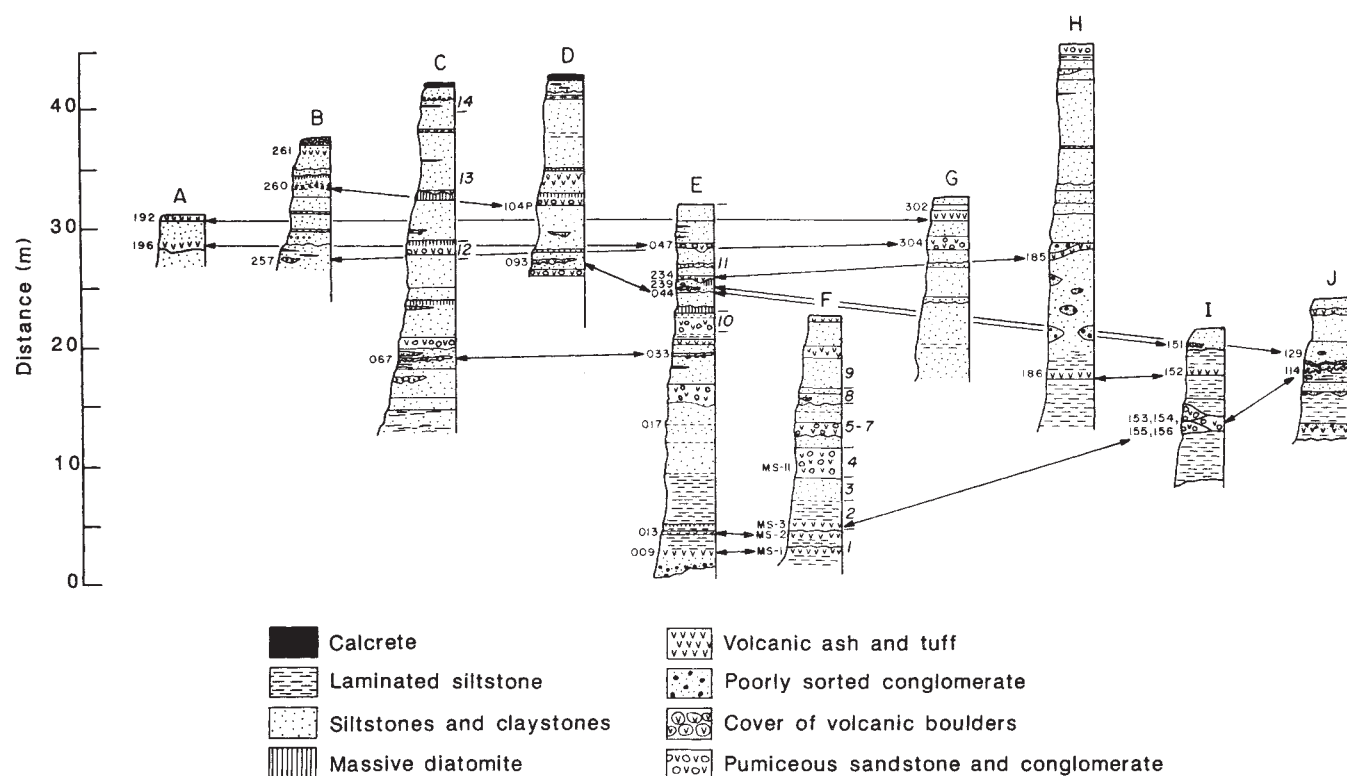


Fig. 2 Stratigraphic sections and correlations of tuffs within the Olorgesailie Formation. Numbers along the right side of columns C, E, and F correspond to members: numbers along the left sides of columns, to sample numbers. Column locations are shown in Fig. 1.

Table 1 Representative compositional data on volcanic glass separates from Olorgesailie and related obsidians

Sample no	Fe ₂ O ₃ *	CaO	K ₂ O	Mn	Nb	Rb	Ti	Y	Zn	Zr	Member
B088(P)	8.4	1.2	4.7	2,270	306	181	4,220	107	221	1,182	14
B087(P)	8.3	1.4	4.8	2,230	358	224	4,210	128	240	1,469	13
B260(P)	8.3	1.1	4.2	2,062	397	253	3,340	145	254	1,700	12
B104(P)	8.0	1.0	4.7	2,120	400	251	3,180	149	263	1,726	12
B100	6.6	0.8	5.2	1,760	440	311	2,480	164	256	2,053	11
B181	6.6	0.7	5.3	1,720	447	319	2,460	165	237	2,106	11
B140	6.9	1.5	5.6	1,370	149	114	5,200	63	192	595	11
B047	6.8	1.4	5.3	1,470	141	107	5,080	57	141	562	11
B093	7.0	1.2	4.8	1,760	270	201	4,340	101	190	1,173	11
B044	6.9	1.2	5.3	1,750	260	171	4,220	98	188	1,116	11
B067	7.9	1.4	4.3	2,060	266	202	3,880	98	207	1,020	9
B031	7.7	1.3	4.8	1,860	275	199	3,770	107	217	1,189	9
B017	6.6	0.5	4.3	1,430	255	249	1,950	171	319	1,675	4
BMS11(P)	6.3	0.4	4.3	1,260	256	255	1,930	168	313	1,621	4
B186	8.2	0.4	4.2	1,600	306	272	1,730	199	366	1,495	2
B152	8.1	0.4	4.5	1,710	301	257	1,700	192	369	1,477	2
B009	6.2	0.6	4.6	1,260	266	245	2,210	159	283	1,480	1
BMS4	6.2	0.6	4.6	970	224	245	2,220	153	280	1,426	1
Olorgesailie											
B302	8.4	1.5		2,570	252	138	4,540	90	211	888	11
Oi Doiyo Nyukie Obsidian											
MER-17	8.5	1.4		2,340	214	129	4,800	89	203	922	
MER-18	8.7	1.4		2,390	225	135	5,060	92	212	971	
Oloronga Obsidian											
HPE-225	8.8	1.4		2,230	216	151	5,390	83	198	899	
HPE-1043	9.0	1.4		2,446	215	147	5,237	88	192	950	

* Total iron as Fe₂O₃; Fe₂O₃, CaO, and K₂O in wt %; others in p.p.m.

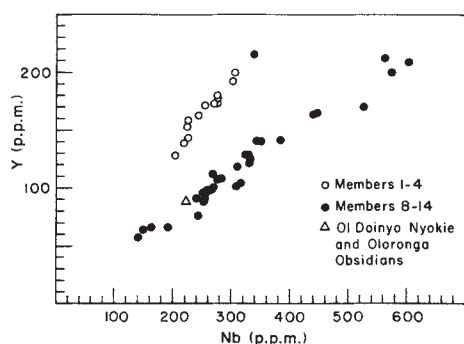


Fig. 3 Plot of niobium against yttrium content of glass separates from the Olorgesailie Formation, and obsidians from Ol Doinyo Nyokie and Oloronga.

composition very similar to that of obsidian from Ol Doinyo Nyokie. The Ol Doinyo Nyokie obsidian is also virtually identical to an obsidian flow from the Oloronga Beds in the Magadi Basin (Table 1). This suggests that strata at Olorgesailie and Oloronga¹⁸ were deposited contemporaneously with volcanism at Ol Doinyo Nyokie. Volcanic rocks from Suswa and Longonot volcanoes appear chemically distinct from both groups represented by the tuffs of the Olorgesailie Formation¹². Tuffs from strata overlying the Olorgesailie Formation are compositionally similar to those in members 8-14 of that formation¹².

The similarity in composition of the upper Olorgesailie Formation to obsidians from Ol Doinyo Nyokie and the Oloronga obsidian is significant because both have been dated at 0.6-0.8 Myr⁶. On this basis alone, one might place members 8-14 of the Olorgesailie Formation at approximately this age. This conclusion is substantiated by potassium argon dates of 0.71 ± 0.02 and 0.74 ± 0.01 Myr on two independent feldspar separates from pumice clasts from member 10 of the Olorgesailie Formation (Table 2). These ages are older than previous determinations of 0.4-0.5 Myr^{16,17}. Palaeomagnetic polarity data suggest that Ol Doinyo Nyokie was active near the Matuyama/Brunhes Chron boundary, as the lowest flows from this volcano are of reversed polarity, whereas younger lavas are normal⁵.

The main archaeological sites at Olorgesailie are situated in members 5-8 of the formation, and there are two disconformities above member 8 in the section. One is at the boundary between member 8 and member 9, the second between members 9 and 10. On the basis of the evidence presented above it is apparent that members 1-9 must predate 0.7 Myr. Isaac¹ has reported that the sediments of member 1 are of normal palaeomagnetic polarity, which together with the regional stratigraphy, suggests that the inception of deposition of the formation occurred no earlier than 0.93 Myr (the older limit of the Jaramillo subchron). With the present data it is not possible to resolve whether these sites fall within the Brunhes Chron, or extend to times as early as the Jaramillo Subchron.

It is likely that members 1-9 of the Olorgesailie Formation lie between 0.93 and 0.70 Myr (possibly between 0.74 and 0.70 Myr), and that members 10-14 lie between 0.7 and 0.6 Myr. The estimate for the younger limit of the latter is taken from

the age reported for a flow from Ol Doinyo Nyokie⁶, the likely source of the ashes found in members 8-14.

We thank P. Gichuki for assistance in the field, G. L. Isaac for providing copies of his unpublished field notes, H. Merrick for collecting samples from Ol Doinyo Nyokie, and H. P. Eugster for samples from the Oloronga Beds. This work was supported by the L. S. B. Leakey Foundation and the National Science Foundation. The National Museums of Kenya provided logistical support. Technical support for the K/Ar dating was provided by T. Davies, R. Maier and A. Sienko. P. Onstott drafted the figures.

Received 27 April; accepted 15 June 1987.

1. Isaac, G. L. *Olorgesailie: archaeological studies of a middle Pleistocene lake basin in Kenya* (University of Chicago Press, Chicago, 1977).
2. Baker, B. H. & Mitchell, J. G. *J. geol. Soc. Lond.* **132**, 467-484 (1976).
3. Baker, B. H., Crossley, R. & Goles, G. G. in *Petrology and Geochemistry of Continental Rifts* (eds Neuman, E.-R. & Ramberg, I. B.) 29-50 (Reidel, Dordrecht, 1978).
4. Baker, B. H. *Geol. Survey Kenya Rep.* **42**, (1958).
5. Baker, B. H. *Bull. Volcanol.* **39**, 420-440 (1976).
6. Fairhead, J. D., Mitchell, J. G. & Williams, L. A. *J. Nature phys. Sci.* **238**, 66-69 (1972).
7. Baker, B. H., Williams, L. A. J., Miller, J. A. & Fitch, F. J. *Tectonophysics* **11**, 191-215 (1971).
8. Shackleton, R. M. in *Geological Background to Fossil Man* (ed. Bishop, W. W.) 171 (Scottish Academic, Edinburgh, 1978).
9. Isaac, G. L. in *Geological Background to Fossil Man* (ed. Bishop, W. W.) 173-206 (Scottish Academic, Edinburgh, 1978).
10. Owen, R. B. & Renault, R. W. in *Paleoecology of Africa* (eds Coetzee, J. A. & Bakker, E. M.) 147-174 (Balkema, Rotterdam, 1981).
11. Brown, F. H. & Cerling, T. E. *Nature* **299**, 212-215 (1982).
12. Bye, B. A. thesis, Univ. Utah (1984).
13. Weaver, S. D. *Bull. Volcanol.* **40**, 209-239 (1976).
14. Weaver, S. D., Seale, J. S. C. & Gibson, I. L. *Contrib. Miner. Petrol.* **36**, 181-194 (1972).
15. Hildreth, W. *Spec. Pap. geol. Soc. Am.* **180**, 43-75 (1979).
16. Evernden, J. F. & Curtis, G. H. *Curr. Anthropol.* **6**, 343-385, (1965).
17. Miller, J. A. in *Background to Evolution in Africa* (eds Bishop, W. W. & Clark, J. D.) 259-272 (University of Chicago Press, Chicago, 1967).
18. Surdam, R. C. & Eugster, H. P. *Bull. Geol. Soc.* **87**, 1739-1752 (1976).

Individuals in an osprey colony discriminate between high and low quality information

Erick Greene

Department of Biology, Princeton University, Princeton, New Jersey 08544, USA

A potential benefit of living in a colony is that animals may gain information about the location of good foraging sites from other colony members¹⁻³. The role of information transfer as a major benefit favouring the evolution of coloniality is, however, very poorly understood⁴. Information transfer has been demonstrated for only a few colonial vertebrate species⁵⁻⁷, but not all colonial species share information⁸⁻¹⁰. Here I report that colonial ospreys (*Pandion haliaetus*, L.) not only transfer information about foraging locations, but that they discriminate among fish species brought back by other colony members, and respond only to schooling prey species. This information significantly reduces the search times needed for informed birds to capture patchily-distributed prey species. This is the first demonstration of a colonial vertebrate discriminating between prey species with different spatial distributions brought back by returning foragers, and selectively using information about location of prey species to increase foraging efficiency.

Ospreys are fish-eating hawks with a cosmopolitan breeding distribution¹¹. They show considerable variation in their social and foraging behaviour, ranging from nesting and foraging solitarily to nesting colonially and foraging in loose groups^{11,12}. Colonial nesting is most frequent in coastal habitats^{13,14}, and before the use of DDT, colonies contained as many as 300 breeding pairs¹⁵. They search for fish by hovering and scanning the water below them. When a fish is chosen they plummet

Table 2 K/Ar analytical data on anorthoclase from Member 10 pumice, Olorgesailie Formation

Lab. no.	K (%)	Weight	⁴⁰ Ar* (10 ⁻¹² moles g)	% ⁴⁰ Ar*	Age and error (Myr ± 1 standard deviation)
ANU 83-9(1)	5.500-5.660	0.5876	6.833	81.6	0.71 ± 0.02
ANU 87-16(1)	5.350-5.335	1.6004	6.856	52.1	0.74 ± 0.01

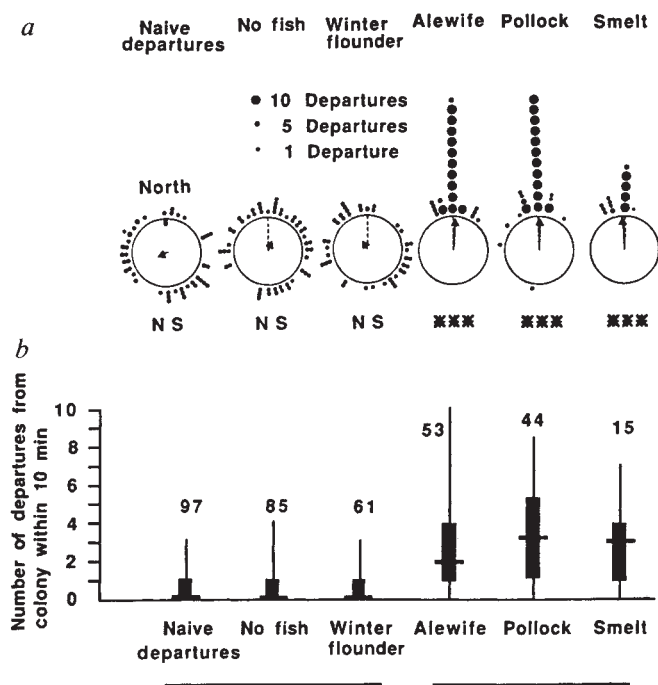


Fig. 1 **a**, The departure directions of ospreys leaving the colony in 10 minute periods after the arrival of another osprey depends upon the species of fish that the returning bird was carrying. Naïve ospreys were defined as birds that departed in randomly-chosen 10 min periods after no osprey had returned to the colony for ≥ 30 min. For naïve birds, departures are relative to true north; for the other 5 classes, the departure distributions are deviations from the incoming direction of the returning osprey (shown as dotted lines). The arrows represent the mean departure angle, and their lengths are proportional to r , the degree of clumping. The distribution of departures for naïve birds does not differ statistically from a circular uniform distribution (Rayleigh test; $r = 0.16$, $Z = 1.31$, $P > 10$). The departure distributions for birds leaving after the arrival of unsuccessful birds (circular V test; $r = 0.15$, $u = 1.4$, $P > 0.10$) or birds with winter flounder (circular V test; $r = 0.14$, $u = 1.2$, $P > 0.10$) do not show significant clumping around 0° . Departures after the arrivals of ospreys with alewife (circular V test; $r = 0.99$, $u = 16.3$, $P < 0.0001$), pollock (circular V test; $r = 0.97$; $u = 16.5$, $P < 0.0001$) or smelt (circular V test; $r = 0.99$, $u = 10.6$, $P < 0.0001$) are highly clumped around 0° . **b**, More ospreys left the colony after a neighbour returned with an alewife, pollock or smelt than at other times. The range, 1st quartile, median and 3rd quartile of the data are shown. The thick lines connect homogeneous sub-groups (pairwise Mann-Whitney U tests; all P values > 0.5 within a sub-group; all P values < 0.001 between sub-groups). The numbers above the ranges are the number of 10 min samples for each category.

feet-first into the water. Successful foragers are conspicuous from a great distance because they fly with the fish held in their feet. Males provide mates and nestlings with most of the food during the breeding season, especially while the chicks are young¹¹. Breeding males usually eat the anterior portion of each fish caught before delivering the remainder to the nestlings. Therefore an osprey rarely resumes foraging within half an hour. Ospreys build large stick nests in exposed locations, which allows a good view of the surrounding areas.

To determine if osprey colonies function as information centres, observations were made at a colony on an island in Cow Bay estuary, Halifax County, Nova Scotia, Canada ($63^\circ 27' \text{W } 44^\circ 37' \text{N}$). The colony contained 11 nests located about 50 m apart. The ospreys fished in Cow Bay, the adjacent ocean, neighbouring estuaries, and inland rivers and lakes up to 5 km away¹⁶. Feeding watches of 4 h duration were conducted from early May to mid-September 1981 (total of 228 observation hours). The observation periods covered all combinations of times of day and variations of the tidal cycle. Observations were made from a bluff that afforded a view of both the colony and foraging areas. Simultaneous observations were also made at other foraging areas frequented by Cow Bay ospreys, so that fishing birds could be followed out of sight of the colony. The locations and species of fish caught were recorded with the time and compass directions of ospreys returning to or leaving the colony. Fish identifications were verified using photographs taken from blinds located near nests. Individual ospreys were identified by their nests and by distinctive moult patterns.

Ospreys nesting in Cow Bay rely upon four fish species which account for more than 90% of their diet by frequency¹⁶: winter flounder (*Pseudopleuronectes americanus*), pollock (*Pollockius virens*), alewife (*Clupeus harengus*) and smelt (*Osmerus mordax*). These fish species vary in their spatial and temporal predictability. Winter flounder are cryptic, bottom-dwelling fish occurring in the estuaries and close inshore during the study period. Although they may enter or leave estuaries in pulses on changing tides¹⁷, winter flounder do not school during this time and tend to be randomly spaced¹⁸. Hence, information about the site of capture of one winter flounder will be a poor predictor of the location of another. During May and early June, alewife and smelt enter inland lakes and rivers to spawn^{19,20}. Schools of these fish gather in the estuary before swimming up river

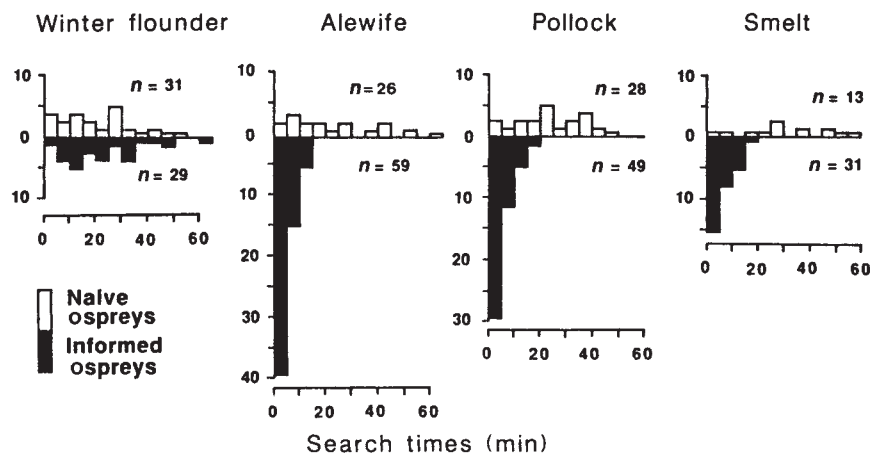
together. During June to mid-July schools of fast-moving juvenile pollock occasionally enter the estuaries²¹. The location of schools or spawning groups of these three species is unpredictable and can change within minutes. In comparison to winter flounder, information about where an osprey caught an alewife, pollock or smelt will be more likely to indicate the location of other fish.

It is clear that this osprey colony functions as an information centre, since ospreys responded to various fish species brought back to the colony (Fig. 1). Non-fishing birds at the colony are more likely to begin fishing following the arrival of a successful forager carrying an alewife, pollock, or smelt than they are after the arrival of an unsuccessful forager or during periods when no ospreys had returned to the colony for at least half an hour (Fig. 1b). Furthermore, these departing birds flew in the direction from which the successful forager had arrived (Fig. 1a). All male ospreys were observed to locate fish schools and also to recruit to schools found by others. This is an important prerequisite for the evolution of information centres: in the absence of other benefits²², consistently successful foragers would gain little by nesting in colonies. In contrast, the arrival of an osprey carrying a winter flounder did not result in increased departure rates (Fig. 1b). The departure directions of ospreys that did leave the colony after the arrival of a winter flounder bore no relationship to the arrival direction of the returning bird (Fig. 1a).

The information transferred at the colony improves the foraging efficiency of colony members. The effect of the information transfer upon foraging efficiency was examined by comparing the search times of naïve and informed ospreys (Fig. 2). Information about where a randomly-distributed prey species (winter flounder) was recently caught does not reduce subsequent search times for that fish species. However, information about the direction where one of the schooling fish species (alewife, pollock, or smelt) was caught greatly reduces search times for those species. Because the hovering flight which ospreys use while fishing is energetically costly, the reduced search times would permit substantial economy.

Information exchange at the colony is both passive and active. Human observers could reliably distinguish the fish species carried by ospreys on the basis of shape and size. Since ospreys have extremely acute eyesight and scrutinize returning birds it

Fig. 2 Search times for naïve versus informed ospreys hunting different fish prey species. Knowing where a winter flounder was recently caught does not reduce an osprey's search time for another winter flounder (Mann-Whitney *U* test; $P = 0.47$), but knowing in which direction a schooling species was recently caught greatly reduces search times for those species (Mann-Whitney *U* tests; all P values < 0.001).



is likely that they discriminate visually among fish species. In addition, male ospreys appear to actively signal the locations of fish schools to other colony members. Successful foragers occasionally performed conspicuous displays while returning to the colony. These birds would call persistently and fly in an undulating fashion towards the colony. The behaviour resembles courtship and threat displays^{11,23}. These displays were typically performed by ospreys that had caught a fish after a long period during which no other colony members had been successful. Furthermore, the displays were observed only after the capture of schooling species (pollock, 9 displays/74 arrivals; alewife, 7 displays/99 arrivals; smelt, 3 displays/64 arrivals; winter flounder, 0 displays/109 arrivals). The display frequencies show a disproportionate association with schooling species; homogeneity test, $G = 10.8$, d.f. = 3, $p < 0.025$). Although these displays were uncommon, the response of other ospreys to this display was dramatic: all of the non-fishing birds at the colony immediately flew towards the displaying bird and began hunting near where the fish had been caught.

Why should ospreys use a costly flight display to share information with other colony members? Male ospreys show high degrees of natal philopatry²⁴, and males at a colony may consequently be close genetic relatives²⁵. Breeding ospreys have been observed feeding young that are not their own, but possibly those of male relatives²⁵. Food-signalling displays may thus be a kin-selected behaviour.

Information exchange at colonies can strongly influence the ospreys' foraging behaviour and prey choice. Although winter flounder were available in Cow Bay throughout the season, their small size and the long search times required to catch them result in low rates of biomass intake¹⁶. In April to June the Cow Bay ospreys ignore these local flounder and hunt alewife, pollock, and smelt, which all give higher rates of food intake. However, it is the informed birds' consistently short search times for pollock, alewife and smelt that make it profitable to ignore local flounder. It is only after these more profitable fish species return to sea that the ospreys begin to hunt flounder¹⁶. Solitary ospreys, without access to information about the locations of fish schools, may have to include winter flounder in their diet much earlier in the season.

Useful information transfer may be limited to areas near osprey colonies. Ospreys do not routinely follow successful foragers back to the foraging site. As the returning bird usually spends at least half an hour eating a portion of its catch. Instead, non-fishing birds at the colony fly in the direction from which the successful osprey arrived. The difficulty of finding distant fish schools solely on the basis of a returning bird's incoming direction may place an upper bound on the distances for which information transfer is effective. Virtually all of the fish were caught within 2 km of the colony, although ospreys recruited to schools of alewife and smelt spawning in streams up to 5 km

inland¹⁶. Another study of an osprey colony located 15 km from the nearest foraging location found that non-fishing birds did not respond to returning foragers²⁶. Instead, individuals were faithful to particular foraging areas over the season and between years.

This study has demonstrated a new level of sensitivity in the ability of a colonial animal to assess the quality of information about the location of potential prey. Moreover, this discrimination allows informed colony members to increase their foraging efficiency through reduced search times for schooling prey species. The fitness consequences of information exchange still need to be evaluated: comparisons of the provisioning rates, prey choice, chick growth rates, and reproductive success between colonial and solitary ospreys nesting in the same area will be revealing. The precise social context and function of the food signalling display also need to be explored.

This work was supported by a Natural Sciences and Engineering Research Council of Canada Undergraduate Scholarship, a Nova Scotia Bird Society Research Grant, and Princeton University. Anne and Greg Greene were invaluable field assistants, and Bill Freedman provided encouragement and support. Peter Austin-Smith of the Nova Scotia Department of Lands and Forests was helpful throughout the course of this study. I thank Charles Brown, Bill Freedman, Peter Grant, Anne Greene, Linda Hamilton, Tony Ives, Bruce Lyon, Yves Prévost, Alan Poole and Will Towne for helpful discussions and thoughtful readings of the manuscript. Special thanks to the Kwindts of Cow Bay, who let us watch ospreys from their bluff.

Received 12 March; accepted 28 July 1987.

1. Ward, P. *Ibis* **107**, 173-214 (1965).
2. Ward, P. & Zahavi, A. *Ibis* **115**, 517-534 (1973).
3. Alexander, R. D. *A. Rev. Ecol. Syst.* **5**, 325-383 (1974).
4. Wittenberger, J. F. & G. L. Hunt, Jr in *Avian Biology* (eds Farner, D. S. & King, J. R.) **8**, 1-78 (Academic, New York, 1985).
5. Bayer, R. D. *Auk* **99**, 31-40 (1982).
6. Brown, C. R. *Science* **234**, 83-85 (1986).
7. Waltz, E. C. *Anim. Behav.* **35**, 48-59 (1987).
8. Loman, J. & Tamm, S. *Am. Nat.* **115**, 285-289 (1980).
9. Andersson, M., Götmark, F. & Wiklund, C. G. *Behav. ecol. Sociobiol.* **9**, 199-202 (1986).
10. Kiis, A. & Möller, A. P. *Anim. Behav.* **34**, 1251-1255 (1986).
11. Cramp, S. in *The Birds of the Western Palearctic*, **2**, (Oxford University Press, 1980).
12. Prévost, Y. A. MSc. Thesis, McGill Univ. Montreal (1977).
13. Spitzer, P. & Poole, A. *Am. Birds* **34**, 234-241 (1980).
14. Greene, E. & Freedman, B. *Can. J. Zool.* **100**, 470-473 (1986).
15. Abbott, C. C. in *The Home Life of the Osprey* (Acar, London, 1911).
16. Greene, E., Greene, A. & Freedman, B. in *Biology and Management of Bald Eagles and Ospreys* (eds Bird, D. M., Seymour, N. R. & Gerrard, J. M.) 257-267 (Harpell, Montreal, 1983).
17. Tyler, A. V. *Fish. Res. Board Canada* **28**, 1727-1732 (1971).
18. Percy, W. G. *Bull. Bingham Oceanographic Coll.* **18**, 1-78 (1962).
19. Messieh, S. *Envir. Biol. Fish.* **2**, 195-210 (1977).
20. Murawski, S. A. & Cole, C. F. *Trans. Am. Fish. Soc.* **107**, 535-542 (1978).
21. Steele, D. H. *J. Fish. Res. Bd. Can.* **29**, 1267-1314 (1963).
22. Weatherhead, P. J. *Am. Nat.* **121**, 237-243 (1983).
23. Harmata, A. *Raptor Res.* **16**, 103-111 (1982).
24. Spitzer, P. thesis, Cornell Univ., Ithaca (1980).
25. Poole, A. *Auk* **99**, 781-784 (1982).
26. Hagen, J. thesis, Northern Carolina State Univ., Raleigh (1986).

Membrane conductance oscillations in astrocytes induced by phorbol ester

Brian A. MacVicar, Stephanie A. Crichton,
Diane M. Burnard & Fred W. Y. Tse

Neuroscience Research Group, Faculty of Medicine,
University of Calgary, 3330 Hospital Drive NW, Calgary,
Alberta, Canada T2N 4N1

Glial cells in the central nervous systems (CNS) have complex functions which are difficult to decipher because of the intimate intertwining of glial cells with neurons. We have therefore developed an essentially neuron-free preparation of CNS astrocytes in the kainic acid lesioned hippocampal slice. With this preparation we have examined the effect of activating protein kinase C in astrocytes with a phorbol ester, TPA (12-*O*-tetradecanoyl-phorbol-13-acetate). In most cells, TPA induced rhythmic oscillations (0.1–3.0 Hz) of membrane potential which were typically 5–10 mV in amplitude and were associated with increases of up to eightfold in input resistance during the depolarizing phase. These large changes in membrane conductance are the first reported observations of endogenously generated conductance changes in astrocytes of the mammalian CNS and they could influence excitability of surrounding neurons, possibly by altering extracellular ion concentrations.

Recent electrophysiological studies have identified several voltage-activated conductances^{1–4} as well as calcium-activated potassium conductances⁵ in cultured glial cells. The physiological significance of these findings, however, is difficult to assess. To address this we have analysed electrophysiological properties of astrocytes in the CA3-CA4 region of the kainic acid (KA) lesioned hippocampal slice, which is essentially neuron-free⁶. Glial cells were identified during intracellular recordings by their high membrane potentials (mean -77.2 ± 8.6 mV, $n = 26$, in 2 mM extracellular K^+), low input resistance (mean 6.7 ± 7.7 M Ω , $n = 14$) and the absence of spiking. Injections of Lucifer Yellow into 30 cells with these characteristics in KA-lesioned slices (D.M.B., S.A.C. and B.A.M., in preparation) showed that these cells had a stellate morphology and extensive dye coupling typical of astrocytes⁷. Intracellular recordings in slices of unlesioned hippocampus demonstrated that normal astrocytes have an average resting potential of -77 mV and an average input resistance of 3.2 M Ω ($n = 15$, W. Walz and B.A.M., in preparation). Therefore, astrocytes in KA-lesioned hippocampal slices had morphological and electrophysiological properties similar to astrocytes in normal hippocampal slices.

In control solution (NaCl, 132 mM; KCl, 2 mM; MgCl₂, 1.3 mM; NaHCO₃, 22 mM; CaCl₂, 2 mM; glucose, 10 mM; pH 7.3) glial membrane potential and input resistance were stable. However 4–15 min after perfusing with TPA (50 nM in 0.01% dimethylsulphoxide (DMSO)), a phorbol ester which activates protein kinase C (ref. 8), 60% of glial cells ($n = 26$) showed rhythmic oscillations of membrane potential with a frequency ranging from 0.1 to 3.0 Hz (Fig. 1). In most experiments input resistance gradually increased by 50–100% following TPA perfusion and then oscillations were observed. Perfusion of 100 nM of an inactive phorbol ester 4 α -PMA (ref. 9, 4 α -phorbol 12-myristate-13-acetate) in 0.01% DMSO did not cause oscillations or any change in input resistance ($n = 4$). The TPA-induced oscillations had an amplitude of 5–10 mV and lasted several seconds or minutes and then ceased for several minutes. The depolarizing phase of the oscillation was always associated with an increase in input resistance as compared to the value at rest and the repolarizing phase was always associated with a decrease. In some cells there was an eightfold difference in input resistance between the depolarized and repolarized phase of the oscillation. Simultaneous intracellular and extracel-

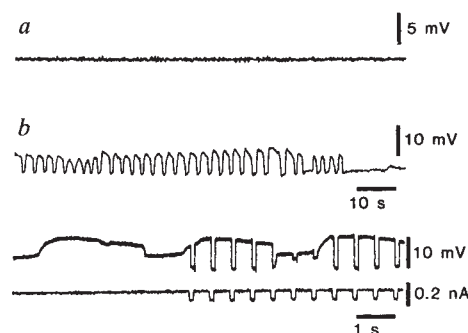


Fig. 1 Phorbol ester induced oscillations in a glial cell. *a*, In control solution, membrane potential (-68 mV) and input resistance (15 M Ω) were stable. *b*, About 5 min after perfusing the slice with TPA (50 nM), rhythmic oscillations of membrane potential and conductance were recorded. They had a frequency of approximately 0.3 Hz (upper trace). Injection of constant current pulses (lower pair of traces) indicated that the input resistance increased to 83 M Ω during the depolarization from 10 M Ω during the hyperpolarization. Immediately before and after the oscillations cell input resistance was stable at 30 M Ω .

Methods. The preparation of hippocampal slices from KA-lesioned rats and the analysis of neuronal lesioning will be described in detail elsewhere (D.M.B., S.A.C. and B.A.M., manuscript in preparation). Briefly, KA (1.0–1.2 μ g) was injected unilaterally or bilaterally into the lateral ventricles. Two to four weeks after injections, slices (500 μ m) were prepared from the hippocampus, incubated in control solution (see text) and transferred individually to a recording chamber. Following experiments, hippocampal slices were fixed in formaldehyde (4%), stained in cresyl violet and acid fuchsin and examined microscopically to confirm the absence of pyramidal neurons in the CA3–CA4 region. KA-induced neuronal degeneration was easily discerned because the CA3 region was visibly stunted. Intracellular and extracellular potentials were recorded with thin-wall glass microelectrodes filled with 2 M potassium acetate (50–60 M Ω), or with 8% Lucifer Yellow (100–120 M Ω) for intracellular staining. Electrodes were visibly positioned in the CA3 region within 50 μ m of each other. Neurons were very rarely impaled in this area indicating that the lesion was extensive.

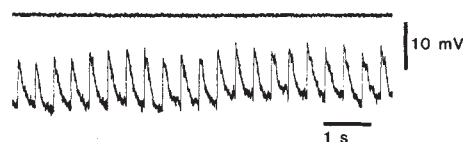


Fig. 2 Simultaneous intracellular (below) and extracellular (above) recordings during TPA-induced oscillations. The transient depolarizations observed intracellularly were associated with decreased conductance, however no changes were observed in the extracellular field potential. These potentials were therefore endogenously generated in the glial cell. The recordings illustrated here were obtained in the presence of Cd²⁺ (1 mM) and apamin (10 nM) which did not prevent the oscillations.

lular recordings indicated that there were no large associated changes in the extracellular field potential (Fig. 2) or in extracellular resistance (not shown). These oscillations were therefore endogenously generated in the glial cells. The frequency of oscillations decreased with membrane depolarization (Fig. 3). In two other cells in which the oscillations were long-lasting and stable, the mean frequency of oscillation decreased by 0.35 Hz and 0.53 Hz per 10 mV depolarization from base frequencies of 2.2 and 2.7 Hz at -85 mV resting membrane potential. The oscillations were not blocked by extracellular Cd²⁺ (1 mM, $n = 4$, Fig. 2), indicating no involvement of calcium

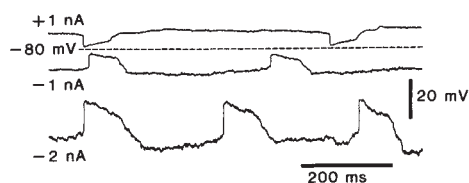


Fig. 3 Effect of membrane potential changes on the oscillation. As membrane potential was altered the frequency of oscillation changed from 1.9 Hz (+1 nA) to 2.7 Hz (-2 nA). The absolute membrane potential during -2 nA current injection was not calibrated because of drifting electrode rectification. Injection of constant current accentuated the amplitude of the oscillations in membrane potential owing to the changes in membrane conductance. Positive deflections during a constant current injection of +1 nA were associated with large decreases in conductance and negative deflections were associated with increases in conductance beyond the steady-state level. During the injection of -1 and -2 nA the waveform reversed so that the large decreases in conductance were associated with negative deflections and the increases in conductance were associated with positive deflections.

entry through the same voltage-activated calcium channels found in cultured astrocytes, which are blocked by Cd^{2+} (ref. 1). These oscillations were still observed when TPA was applied in the presence of 10 mM tetraethylammonium (TEA) ($n=4$) or 10 nM apamin ($n=4$). These concentrations are effective in blocking some types of calcium-activated potassium channels^{5,10}.

We conclude that these oscillations in membrane potential and input resistance result from protein kinase C activation in glial cells by TPA because the inactive phorbol ester did not evoke such oscillations. How might protein kinase C activation cause oscillations? Protein kinase C activation has been observed to cause oscillating intracellular calcium concentrations in activated oocytes¹¹. It is possible that protein kinase C activation also causes oscillating levels of intracellular free calcium in astrocytes. Calcium could be released from internal stores or enter through a type of calcium channel that is not blocked by Cd^{2+} . Transient hyperpolarizations, similar to the repolarizing component of these oscillations, have been reported in fibroblasts¹² and in dorsal root ganglion (DRG) neurons after caffeine exposure¹³ and are due to a calcium-activated potassium current evoked by intracellular calcium release. Cultured astrocytes have a calcium-activated potassium channel⁵ and the changes in input resistance during TPA-induced oscillations could also be due to changing membrane potassium permeability activated by calcium. Possibly a type of calcium-activated potassium conductance is involved that is not as sensitive to extracellular TEA or apamin. Alternatively, another intracellular messenger system may be involved, such as cyclic AMP, which could also alter potassium permeability.

What are the possible consequences of such oscillations and could they occur in normal tissue? These oscillations have been observed occasionally (4 out of 74 cells) in glial cells from normal hippocampi (W. Walz and B.A.M., in preparation) and also in glial cells in control solution in KA-lesioned hippocampal slices (2 out of 35 cells). It is therefore not necessary to apply TPA to observe these oscillations in normal and reactive astrocytes. Astrocytes in primary culture have protein kinase C which is activated by phorbol esters¹⁴ and they have receptors for transmitters which increase phosphatidylinositol turnover¹⁵⁻¹⁸ and could activate protein kinase C. Therefore, it is possible that protein kinase C in astrocytes in the intact CNS could be activated by neurotransmitters which could then elicit the oscillations reported here. These fluctuations in glial membrane conductance could profoundly affect homeostasis of the ionic environment of adjacent neurons.

Supported by the MRC of Canada and The Alfred P. Sloan

Foundation. D.M.B. and F.W.Y.T. are Alberta Heritage Foundation for Medical Research Postdoctoral Fellows and B.A.M. is an Alberta Heritage Foundation for Medical Research Scholar and a Sloan Fellow.

Received 13 March; accepted 13 July 1987.

- MacVicar, B. A. *Science* **226**, 1345-1347 (1984).
- Bevan, S., Chiu, S. Y., Gray, P. T. A. & Ritchie, J. M. *Proc. R. Soc. B* **225**, 299-313 (1985).
- Bevan, S. & Raff, M. *Nature* **315**, 229-232 (1985).
- Gray, P. T. A. & Ritchie, J. M. *Proc. R. Soc. B* **228**, 267-288 (1986).
- Quandt, F. N. & MacVicar, B. A. *Neuroscience* **19**, 29-41 (1986).
- Ashwood, T. J., Lancaster, B. & Wheal, H. V. *Exp. Brain Res.* **62**, 189-198 (1986).
- Gutnick, M. J., Connors, B. W. & Ransom, B. R. *Brain Res.* **213**, 486-492 (1981).
- Nishizuka, Y. *Nature* **308**, 693-698 (1984).
- Kaczmarek, L. K. *J. exp. Biol.* **124**, 375-392 (1986).
- Romey, G., Hugues, M., Schmid-Antomarchi, H. & Lazdunski, M. *J. Physiol., Paris* **79**, 259-264 (1984).
- Cuthbertson, K. S. R. & Cobbold, P. H. *Nature* **316**, 541-542 (1985).
- Henkart, M. P. & Nelson, P. G. *J. gen. Physiol.* **73**, 655-673 (1979).
- Kuba, K. *J. Physiol.* **298**, 251-269 (1980).
- Mobley, P. L., Scott, S. L. & Cruz, E. G. *Brain Res.* **398**, 366-369 (1986).
- Murphy, S., Pearce, B. & Morrow, C. *Brain Res.* **364**, 177-180 (1986).
- Pearce, B., Cambray-Deakin, M., Morrow, C., Grimble, J. & Murphy, S. *J. Neurochem.* **45**, 1534-1540 (1985).
- Torrens, Y. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 9216-9220 (1986).
- Fisher, S. K. & Agranoff, B. W. *J. Neurochem.* **48**, 999-1017 (1987).

Direct measurement of proton transfer rates to a group controlling the dihydropyridine-sensitive Ca^{2+} channel

Blaise Prod'homme, Daniela Pietrobon & Peter Hess

Department of Physiology and Biophysics, Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA

Protons participate in most biologically important reactions, as substrates, products, cofactors and modulators, and proton transport is an essential step in energy transduction^{1,2}. The dynamics of protonation reactions have been studied extensively in solution and in model systems involving lipid-water interfaces³⁻⁶, but have never been resolved at the timescale of the elementary molecular event. Here we show that, under appropriate conditions, binding and unbinding reactions of single protons and deuterium ions to a single site on the L-type calcium channel can be resolved and the protonation and deprotonation rates quantified. The protonation rate constant considerably exceeds previous estimates obtained in simpler systems. The functional consequences of channel protonation is a threefold reduction of the channel conductance, independent of the applied voltage. The data are consistent with the presence of a single protonatable group with pK in the physiological pH range, close to the external mouth of the channel. The two conductance levels of the open channel might be explained by greatly differing local potentials associated with the protonated and deprotonated state of the group.

Transmembrane Ca^{2+} currents are very sensitive to changes in extracellular pH (pH_o ; refs 7-10). However, we found that under standard conditions for single Ca^{2+} channel recordings, with 110 mM Ba^{2+} as the charge carrier¹¹⁻¹³, the unitary current

Table 1 Rate constants for the association and dissociation of H^+ and D^+ ions with a site on the L-type Ca^{2+} channel

Ion	<i>n</i>	$k_{on} (\text{M}^{-1} \text{s}^{-1})^*$	$k_{off} (\text{s}^{-1})$	<i>pK</i>
D^+	6	$3.80 \pm 0.28 \times 10^{11}$	$0.68 \pm 0.03 \times 10^4$	7.75
H^+	4	$4.09 \pm 0.15 \times 10^{11}$	$1.67 \pm 0.05 \times 10^4$	7.40

*Values for k_{on} were pooled after conversion of the measured first-order rate constant at each pH (or pD) to a second-order rate constant by division by the H^+ (or D^+) concentration (*n*, number of measurements).

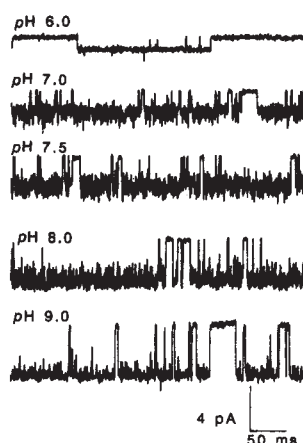


Fig. 1 Cell-attached patch recordings of single L-type Ca^{2+} channel currents carried by Na^+ ions at the pH_o values indicated above each sweep. The traces were obtained during depolarizations to -40 mV from holding potentials ranging from -70 to -110 mV. This voltage protocol is adequate for activation of L-type Ca^{2+} channels in the absence of divalent ions¹⁵. The currents were sampled at 5 kHz and filtered at 1 kHz (-3 dB, 8-pole Bessel filter). **Methods.** Single guinea pig ventricular myocytes were freshly dissociated by enzymatic dispersion¹⁵. Cell-attached patch clamp recordings³⁶ were made at room temperature (21°C) with depolarizing pulses applied every 3–4 s. Patch pipettes contained 150 mM NaCl and 5 mM EDTA. The pH buffer (5 mM) was TAPS ($\text{pK} = 8.4$) for the pipette solution at pH 9, HEPES ($\text{pK} = 7.55$) for the solutions at pH 7–8 and MES ($\text{pK} = 6.15$) for the solution at pH 6. The cell membrane potential outside the patch was zeroed¹⁵ by an external solution containing (in mM): potassium aspartate 140, EGTA 10, HEPES 10, MgCl_2 20. The dihydropyridine Ca^{2+} channel agonist (+)-(S)-202-791 (gift from Dr Hof, Sandoz Co.) was added at $1\text{ }\mu\text{M}$ to the external solution to promote long-lasting channel openings^{18–20}.

amplitude (i) of the open L-type Ca^{2+} channel¹⁴ was little changed between pH 9 and pH 6 (results not shown). To unmask protonation sites which might also bind divalent ions and would therefore be saturated by the very high Ba^{2+} concentration, we chose to record single Ca^{2+} -channel currents in the absence of divalent ions with Na^+ as the charge carrier^{15,16}. Single L-type Ca^{2+} channels in guinea pig ventricular myocytes¹⁴ were identified by their unitary conductance to Na^+ , voltage dependence, gating kinetics and sensitivity to dihydropyridine Ca^{2+} -channel agonists¹⁵, like (+)-(S)-202-791 (ref. 17). This drug was used in all experiments to promote the appearance of long-lasting channel openings^{18–20}. Under these recording conditions, H^+ ions dramatically change the single channel currents (Fig. 1). Fast flickery block reduces i from the fully open level at pH 9 to a much reduced but non-zero level at pH 6.

The underlying elementary current steps could be revealed by substitution of D_2O for H_2O , because the flickery transitions were slowed just enough to bring them within the bandwidth (5 kHz) of our recording system. Channel openings are interrupted by brief transitions between the fully open level of ~ 4.4 pA and a partially blocked level of ~ 1.5 pA (Fig. 2). With increasing D^+ concentrations these incomplete blocking events occur more frequently. The different current levels are shown clearly in the amplitude histograms associated with each pD in Fig. 2. The fully open level predominates at pD 8.55, whereas the appearance of the blocked level increases with decreasing pD .

The simplest and most likely explanation for the current traces and amplitude histograms in Fig. 2 is that individual blocking transitions result from binding and unbinding of individual D^+ ions to a single site on the channel. In this case the fraction of time an open channel spends in the blocked state should increase

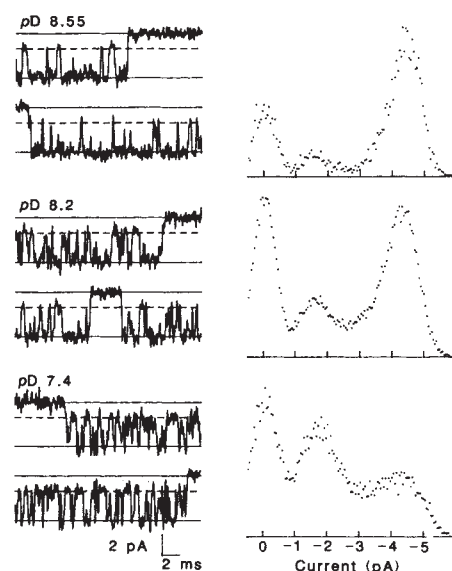


Fig. 2 Individual deuteronation events resolved at three pD values. Current traces (left panels) and amplitude histograms at the indicated values of pD . Note expanded timescale for the currents. The traces were sampled at 25 kHz and filtered at 5 kHz. Voltage protocol as for Fig. 1. The three lines in each trace represent the zero current level of the closed channel (upper solid line), the blocked level at -1.5 pA (broken line) and the fully open level at -4.4 pA. They correspond to the three peaks in the amplitude histograms.

Methods. Bath solution as in Fig. 1. Pipette solutions: NaCl (150 mM), EDTA (2–5 mM) and pH buffer (2.5 mM TAPS for $\text{pD} > 8$, 5 mM HEPES for $\text{pD} < 8$) were dissolved in D_2O (99.8%, Aldrich). The pD was measured with a pH electrode calibrated in pH buffers as $\text{pD} = \text{meter reading} + 0.4$ (ref. 37). Amplitude histograms were obtained from segments of open channel. A variable amount of closed channel was included for each patch, such that the height of the peaks at 0 pA is arbitrary. Scaling of each histogram is arbitrary, but as expected if each D^+ ion causes one blocking transition, the ratio of the peak at the blocked level to the peak at the fully open level increases linearly with the D^+ concentration. The ratios are 0.16 (pD 8.55), 0.38 (pD 8.2) and 2.1 (pD 7.4) and they represent the ratio of the (first order) on- and off-rate constants for D^+ ions at each pD .

linearly with the concentration of D^+ ions. This was indeed the case and can be seen most directly by comparison of the ratio of the amplitudes of the peaks associated with the fully open and the blocked level at each pD (see Fig. 2 legend). Measurements of the mean duration of the blocked intervals and the intervals between successive blocking events should then yield direct estimates for the rate constants underlying the binding and unbinding reaction^{21,22}. An example of such an analysis is shown in Fig. 3. Distributions of the blocked and unblocked intervals are shown at pD 8.65 and pD 8.2. The distributions are well fitted by a single exponential, as expected for a process with a single blocked (bound) and unblocked state. Also as expected, the mean duration of the deuterated state is independent of the pD , whereas the mean interval between successive deuteronation events (mean unblocked time) shortens with increasing D^+ concentration. The mean values for the rate constants for the binding and unbinding reaction of D^+ ions are given in Table 1.

The experiments with D_2O established the elementary 'pattern' of block and permitted direct estimation of deuteronation-deuteronation rate constants. To extract the corresponding protonation-deprotonation rate constants, we used the information contained in the amplitude histograms because the elementary blocking events were too short to be measured directly (Fig. 4). We simulated current fluctuations between two states of

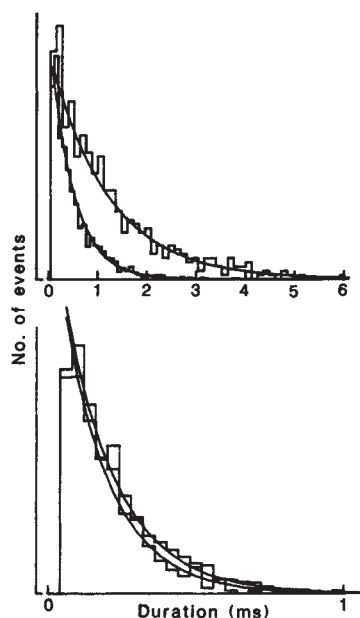


Fig. 3 Kinetic analysis of the D^+ -channel interaction. Superimposed distributions of intervals between blocking events (top) and of blocked times (bottom) at pD 8.65 and pD 8.2. Each distribution is fitted with a single exponential (non-linear χ^2 minimization routine). The time constants (τ) for the best fits are, for the upper distribution, 1.20 ms at pD 8.65 and 0.48 ms at pD 8.2; for the lower distribution, 0.16 ms at pD 8.65 and 0.18 ms at pD 8.2. Blocked and unblocked intervals were obtained by an idealization procedure based on a discriminator placed midway between the fully open and the blocked level³⁸. The effect of missed short events was evaluated by the method introduced by Neher³⁹. For the example shown, the corrected time constants at pD 8.65 and 8.2 are 0.90 and 0.36 ms for the intervals between blockings and 0.14 and 0.15 ms for the blocking events. This correction was used for the calculation of the rate constants given in Table 1.

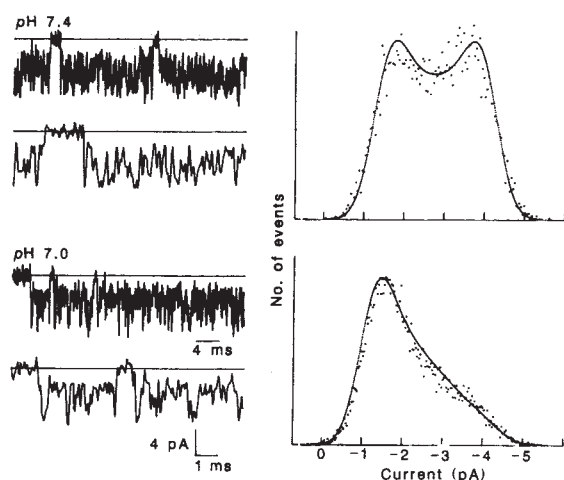


Fig. 4 Quantitative analysis of protonation-deprotonation rate constants. Individual traces at pH 7.4 and 7.0 (left panels) and corresponding amplitude histograms (right panels). Solutions and voltage protocol as in Fig. 1. Current records were digitized at 25 kHz and filtered at 5 kHz. Note different timescales for upper and lower traces. Segments of complete channel closings were excluded from the amplitude histograms, which therefore represent only the current fluctuations between the open and blocked levels. Solid curves superimposed on the data points, simulations (see text) obtained with the rate constants as follows: top, $k_{on} = 16,000 s^{-1}$, $k_{off} = 16,000 s^{-1}$; below, $k_{on} = 40,000 s^{-1}$, $k_{off} = 16,700 s^{-1}$. The two kinetic parameters for the simulations were changed in increments of no more than $\pm 5\%$ and the best fit determined by eye.

appropriate amplitudes. The resulting signal was passed through the same electronic filter as the one used for the experiments. The amplitude histogram of the simulated data was then convolved with a gaussian distribution of the same width as that of the closed channel (to include background noise) and superimposed on the experimental data. Simulations were run with different rate constants until a satisfactory fit to the data could be obtained (Fig. 4). This straightforward but rather slow procedure was preferred over previously published analytical methods to fit amplitude histograms of filtered two-state processes²³, because it eliminates the need for simplifying assumptions about the filter response²³.

Spectral analysis of the open-channel current fluctuations revealed a single lorentzian component, consistent with the assumption that the proton-induced current fluctuations were the only source of extra noise in our records. Moreover, when the variance obtained from the power spectrum was used to calculate the rate constants, with correction for the limited frequency response²⁴, the numbers very closely matched those obtained with the amplitude histogram method (Fig. 4) listed in Table 1.

Several points are worth noting. First, the protonation rate constants are $\sim 1,000$ times higher than the diffusion-limited 'on' rate constants of other cations like Na^+ or Ca^{2+} to organic molecules^{25,26}, soluble proteins²⁷ or ion channels²⁸ including the L-type Ca^{2+} channel²². This finding is consistent with the much higher mobility of protons in solution, which is believed to arise from the very rapid exchange of protons between H_3O^+ and H_2O molecules²⁹.

Second, the protonation rates are extremely high. They exceed the diffusion-controlled reaction rate constants of protons with bases in aqueous solution (10^{10} – $10^{11} (M s)^{-1}$, see ref. 3) and are also 40 times faster than the rate constants estimated for the protonation of acid groups on micelles⁴. Factors contributing to the very rapid protonation rate constants must include electrostatic interactions between proton and base. Such interactions could be considerably stronger at the surface of a protein than in free solution, if a negatively charged proton acceptor is located in a recess like the channel vestibule, because the low dielectric of the protein surrounding the vestibule would funnel the electric field lines³⁰ and thus greatly extend the length over which the electrostatic attraction in the aqueous phase is felt by an incoming proton. Similarly, the presence of extra negative charge (on amino-acid or sugar residues with $pKs < 6$) close to the proton acceptor would speed up its protonation. Such extra negative charges could be thought of as effectively increasing the capture radius for a successful collision between proton and base⁵. Whatever the underlying mechanism, the 'superfast' protonation dynamics must originate from the protein microenvironment. Not surprisingly, the apparent pK value for protons of 7.5 does not correspond to any pK value of an amino-acid side chain in water, thus making it impossible for us to use the pK value for the identification of the amino-acid residue involved.

Third, although the binding rate constants for H^+ and D^+ ions are similar, the rate constant for unbinding of D^+ is slower by about a factor of 2.5. This is a direct demonstration that deuterium bonds are more stable than proton bonds^{31,32}. This difference is reflected in the higher pK value (see Table 1) with deuterium and probably explains the fact that equilibrium binding constants for deuterium are generally higher than the corresponding ones for protons.

The elementary current steps associated with protonation and deprotonation of the L-type Ca^{2+} channel point to a novel mechanism of functional modification. A simple blocking mechanism is unlikely, because in all cases described so far, the blocking ion completely shuts off the channel for the duration of its intra-channel dwell time (see refs 22, 33 and 34). The rapid protonation rate suggesting diffusion-controlled access and the fact that the measured on and off rates both with H^+ and D^+ ions were independent of the membrane potential over

the voltage range studied (-20 to -70 mV, unpublished results) represent strong evidence for an external location of the proton binding site, outside the part of the channel which experiences the transmembrane electric field. The effect of protonation could be to trigger a conformational change in the protein which would reduce the channel conductance. Although we cannot rule out such a mechanism at the moment, we prefer a model in which the ionizable site is near the external entrance of the channel. The channel conductance varies with the protonation state of the site, because its protonation-dependent charge sets up a local potential which attracts counterions into the pore and repels cations. Thus, at equal bulk concentrations of Na^+ , the Na^+ concentration at the channel entrance should be higher at pH 9 than at pH 6. Consistent with this scheme, we found that the channel conductance for Na^+ saturated at much lower bulk Na^+ concentrations at pH 9 than at pH 6 (data not shown). Effects of local surface charge on channel permeation have long been hypothesized (see ref. 34) and recently a formal model was introduced³⁰ for the case of a charge localized in the vestibule of a channel. Our results are, however, the first ones at the unitary level in direct support for such a mechanism.

It is not clear to what extent a similar mechanism of proton-channel interaction might be important in the permeation mechanism for other ion channels. The results are particularly clear for the L-type Ca^{2+} channel because there is only one functionally important acid group with a pK between pH 9 and 6, and the off rate is just slow enough to lead to detectable unitary fluctuations. Multiple sites with lower affinity would be predicted to cause merely an increase in the open channel noise. Such noise has been detected in other channels³⁵ but has not so far been correlated with protons.

This work was supported by the USPHS, the Milton Fund, and fellowships from the Muscular Dystrophy Association (B.P.) and the European Molecular Biology Organization (D.P.).

Received 1 June; accepted 16 July 1987.

- Mitchell, P. *Nature* **191**, 144-148 (1961).
- Boyer, P. D. *et al. A. Rev. Biochem.* **46**, 955-1026 (1977).
- Bell, R. P. *The Proton in Chemistry* (Cornell University Press, Ithaca, 1973).
- Gutman, M., Nachliel, E., Gershon, E. & Giniger, R. *Eur. J. Biochem.* **134**, 63-69 (1983).
- Gutman, M. & Nachliel, E. *Biochemistry* **24**, 2941-2946 (1985).
- Kasianowicz, J., Benz, R. & McLaughlin, S. *J. Membrane Biol.* **95**, 73-89 (1987).
- Iijima, T., Ciani, S. & Hagiwara, S. *Proc. natn. Acad. Sci. U.S.A.* **83**, 654-658 (1986).
- Irisawa, H. & Sato, R. *Circulation Res.* **59**, 348-355 (1986).
- Krafte, D. S. & Kass, R. S. *Biophys. J.* **49**, 433a (1986).
- Konnerth, A., Lux, H. D. & Morad, M. *J. Physiol., Lond.* **386**, 603-633 (1987).
- Fenwick, E. M., Marty, A. & Neher, E. *J. Physiol., Lond.* **331**, 599-635 (1982).
- Reuter, H., Stevens, C. F., Tsien, R. W. & Yellen, G. *Nature* **297**, 501-504 (1982).
- Cavalié, A., Ochi, R., Peltzer, D. & Trautwein, W. *Pflügers Arch. ges. Physiol.* **398**, 284-297 (1983).
- Nilius, B., Hess, P., Lansman, J. B. & Tsien, R. W. *Nature* **316**, 443-446 (1985).
- Hess, P., Lansman, J. B. & Tsien, R. W. *J. gen. Physiol.* **88**, 293-319 (1986).
- Matsuda, H. *Pflügers Arch. ges. Physiol.* **407**, 465-475 (1986).
- Hof, R. P., Ruegg, U. T., Hof, A. & Vogel, A. *J. Cardiovasc. Pharmac.* **7**, 689-693 (1985).
- Kokubun, S. & Reuter, H. *Proc. natn. Acad. Sci. U.S.A.* **81**, 4824-4827 (1984).
- Hess, P., Lansman, J. B. & Tsien, R. W. *Nature* **311**, 538-544 (1984).
- Kokubun, S., Prod'homme, B., Becker, C., Porzig, H. & Reuter, H. *Molec. Pharmac.* **30**, 571-584 (1986).
- Neher, E. & Steinbach, J. H. *J. Physiol., Lond.* **277**, 153-176 (1978).
- Lansman, J. B., Hess, P. & Tsien, R. W. *J. gen. Physiol.* **88**, 321-347 (1986).
- Yellen, G. *J. gen. Physiol.* **84**, 157-186 (1984).
- Ogden, D. C. & Colquhoun, D. *Proc. R. Soc. B225*, 329-355 (1985).
- Bayley, P., Ahlstrom, P. Martin, S. R. & Forsen, S. *Biochem. biophys. Res. Commun.* **120**, 185-191 (1984).
- Diebler, H., Eigen, M., Ilgenfritz, G., Maas, G. & Winkler, R. *Pure appl. Chem.* **20**, 93-115 (1969).
- Potter, J. D., Johnson, J. D. & Mandel, F. *Fedn Proc.* **37**, 1608 (1978).
- Andersen, O. S. *Biophys. J.* **41**, 147-165 (1983).
- Eisenberg, D. & Kauzmann, W. *The Structure and Properties of Water* (Oxford University Press, New York, 1969).
- Dani, J. *Biophys. J.* **49**, 607-618 (1986).
- Katz, J. J. *Am. Sci.* **48**, 544-580 (1960).
- Jencks, W. P. *Catalysis in Chemistry and Enzymology* (McGraw-Hill, New York, 1969).
- Fukushima, Y. *J. Physiol., Lond.* **331**, 311-331 (1982).
- Hille, B. *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, Massachusetts, 1984).
- Sigworth, F. J. *Biophys. J.* **47**, 709-720 (1985).
- Hamill, O., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. *Pflügers Arch. ges. Physiol.* **391**, 85-100 (1981).
- Perrin, D. D. & Dempsey, B. *Buffers for pH and Metal Ion Control* (Chapman and Hall, London, 1974).
- Colquhoun, D. & Sigworth, F. J. in *Single Channel Recordings* (eds Sakmann, B. & Neher, E.) 191-263 (Plenum, London, 1983).
- Neher, E. *J. Physiol., Lond.* **339**, 663-678 (1983).

Genetic linkage of bilateral acoustic neurofibromatosis to a DNA marker on chromosome 22

Guy A. Rouleau*, Wladimir Wertelecki†, Jonathan L. Haines‡, Wendy J. Hobbs*, James A. Trofatter‡, Bernd R. Seizinger*, Robert L. Martuza§, Duane W. Superneau†, P. Michael Conneally‡ & James F. Gusella*

* Neurogenetics laboratory, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA

† Department of Medical Genetics, University of South Alabama, USA Medical Center (MCSP 1070), Mobile, Alabama 36617, USA

‡ Department of Medical Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46223, USA

§ Department of Surgery, Neurosurgery Service, Massachusetts General Hospital, and Harvard Medical School, Boston, Massachusetts 02114, USA

Bilateral acoustic neurofibromatosis (BANF) is a severe autosomal dominant disorder involving development of multiple tumours of the nervous system including meningiomas, gliomas, neurofibromas and particularly bilateral acoustic neuromas. We have used genetic linkage analysis with DNA markers to establish that the defective gene causing BANF is on chromosome 22, and is therefore distinct from the gene for the von Recklinghausen form of neurofibromatosis, which maps to chromosome 17. Linked DNA markers will be particularly valuable in BANF, facilitating early detection of tumours and thereby permitting more effective surgical intervention. In view of the reported loss of genes on chromosome 22 in meningiomas and acoustic neuromas, the genetic localization of the primary BANF defect strongly supports the concept that the disease locus encodes a 'tumour suppressor' gene. Isolation of this gene should provide insights into the pathogenesis of acoustic neuromas and other nervous system tumours, as well as into the control of proliferation and differentiation of neural crest cells.

Neurofibromatosis, one of the most common mendelian disorders affecting the nervous system¹, has two clinically distinct forms: von Recklinghausen neurofibromatosis (VRNF)² and bilateral acoustic neurofibromatosis (BANF)³. Both are transmitted as autosomal dominant defects causing predisposition to various tumours of the nervous system, particularly benign Schwann cell tumours. In VRNF, Schwann cell tumours are typically confined to the skin, peripheral nerves and spinal nerve roots; central nervous system tumours, other than benign optic

Table 1 Lod scores for linkage of BANF to D22S1

Marker locus	Recombination fraction (θ)						
	0.00	0.01	0.05	0.10	0.20	0.30	0.40
D22S1	3.42	3.37	3.13	2.48	2.11	1.35	0.59

Peak lod score ($\hat{z}$) was 3.42 at $\hat{\theta} = 0.0$.

gliomas, are rare. By contrast, bilateral acoustic neuromas, Schwann cell tumours of the vestibular branch of the eighth cranial nerve, are the hallmark of BANF; other manifestations include multiple cranial and spinal meningiomas and nerve root neurofibromas as well as gliomas of the brain stem and the spinal cord. These central nervous system tumours often lead to deafness, balance disorder and paralysis necessitating multiple surgical procedures in the teens or in early adulthood. Though VRNF is more common (rate of occurrence 1 in 3,000), the neurological morbidity of BANF (1 in 100,000) can be more

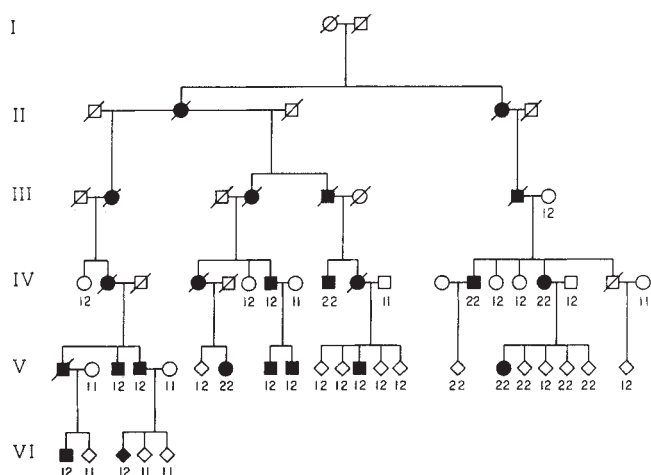


Fig. 1 Bilateral acoustic neurofibromatosis pedigree. The following symbols are used: $\square$, male; $\circ$, female; $\triangle$, sex not specified so as to preserve confidentiality of individuals at risk; solid symbol, affected with BANF; slashed symbol, deceased. Diagnosis of BANF was established according to the following criteria; first computerized imaging (computerized tomography (CT) or magnetic resonance imaging (MRI)) evidence of bilateral internal auditory canal masses, or second, one of the following where a parent, sibling or child had documented BANF: (1) unilateral CT or MRI evidence of an internal auditory canal mass; (2) one of the following at any site—meningioma, glioma, ependymoma or Schwann cell tumour. Carriers of the BANF gene generally develop symptoms by 25 years of age, the mean age of onset of symptoms being 20.4 years (ref. 3). Penetrance is reported to be over 95% (ref. 3). We have documented no cases of non-penetrance in this family. Apparent genotypes for *D22S1* are indicated on the pedigree and were determined by Southern blotting using genomic DNA prepared from blood and/or lymphoblast lines as previously described¹⁸.

devastating and is exacerbated by the difficulty of early diagnosis. A major question which has remained unresolved is whether the two forms of neurofibromatosis represent allelic variants or distinct genetic loci.

Recent studies of BANF-associated tumours have implicated chromosome 22 in the pathogenesis of this disease⁴. Restriction fragment length polymorphisms (RFLPs) have been used to show that loss of alleles on chromosome 22 is a frequent event both in unilateral acoustic neuroma, a common sporadic tumour, and in the bilateral acoustic neuromas characteristic of BANF. A similar loss of chromosome-22 alleles has been observed in sporadic and BANF-associated meningiomas^{4,5} and in neurofibromas from BANF patients⁴. Although these studies demonstrate an association between loss of alleles on chromosome 22 and tumour formation, they do not establish whether this chromosomal mechanism reflects the primary cause or simply a secondary consequence of tumorigenesis. To resolve this critical issue, we have performed genetic linkage analysis with DNA markers to determine whether the primary inherited defect in BANF is located on chromosome 22.

Pedigree and clinical data were compiled from a single large BANF kindred ideal for genetic linkage studies (Fig. 1)⁶. The diagnosis of BANF was established according to the criteria outlined in Fig. 1. None of the affected individuals met the NIH criteria for VRNF (ref. 1). We typed BANF family members for *D22S1*, a marker frequently reduced to hemizygosity in BANF-related tumours. *D22S1*, detected by the probe pMS3-18, displays an apparent single-site *Bgl*III RFLP with allelic fragments of 9.5 kilobases (kb) (allele 1) and 6.5 kb (allele 2) having frequencies of 0.81 and 0.19, respectively^{8,9}. The data was analysed with the computer program LIPED (ref. 10), assuming penetrance of 0.95 and with an age of onset correction based on a mean of 20 years (refs 3 and 11), to calculate a lod score

(z) for linkage of the marker to the defect assuming various recombination fractions (θ) (Table 1). The lod score had a maximum ($\hat{z} = +3.42$) at $\theta = 0.0$ which exceeds the critical value (+3.0) accepted as proof of linkage, and indicates the absence of any detectable crossovers between the marker and the disease locus. The one-lod-unit confidence interval spans ~10% recombination on either side of the marker locus. These data clearly establish that the defect causing BANF is on the long arm of chromosome 22, in the immediate vicinity of *D22S1*. This DNA marker has previously been genetically mapped proximal to *SIS*, in 22q12.3–22q13.1 (refs 12 and 13). The BANF gene therefore maps in the centre of the chromosome-22 long arm, probably in the region 22q11.1–22q13.1 shown in Fig. 2. Further linkage analyses with the markers *IGLV* and *SIS*, which have been physically mapped with more precision than *D22S1*, should clarify the exact position of the defect.

Knudson proposed that retinoblastomas develop as a consequence of two separate genetic events possibly involving loss-of-function mutations at a specific locus¹⁴. In familial cases, predisposition to tumour formation is transmitted in autosomal dominant fashion, with tumours resulting from the unmasking of the inherited mutant allele by somatic mutation affecting the normal homologue. Molecular genetic evidence has provided support for this model in several different tumour types¹⁵. The demonstration that the genetic defect underlying BANF is in the same region lost from the tumours characteristic of the disorder strongly supports the view that the mechanism proposed by Knudson also applies to acoustic neuromas and BANF, possibly by altering a gene critical in controlling Schwann cell proliferation and differentiation.

BANF is similar to, but clinically distinct from, VRNF (refs 2 and 3), which is caused by a defect recently mapped to chromosome 17 (ref. 16). We have tested for linkage of the BANF gene to DNA markers on chromosome 17 with negative results (G.A.R. and J.F.G., unpublished data). Furthermore, linkage of VRNF to *D22S1* has also been excluded (ref. 17, and G.A.R. and J.F.G., unpublished data). Thus, despite the formation of Schwann cell tumours in both disorders, BANF and VRNF are not caused by allelic mutations, but by defects in discrete loci.

We have studied only one family with BANF and so have not addressed the question of nonallelic genetic heterogeneity, but the strong association of loss of alleles from this chromosome and acoustic neuromas⁴ makes it unlikely. When more precise recombination estimates are available, markers linked to BANF should be applicable to presymptomatic or prenatal diagnosis

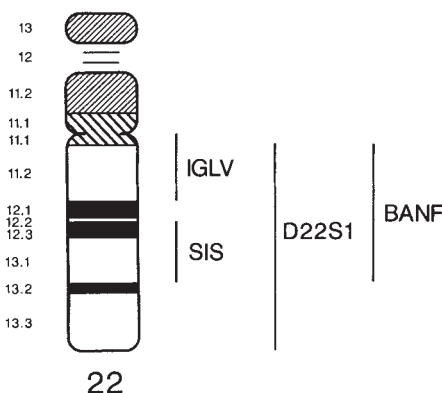


Fig. 2 Location of DNA markers on chromosome 22. The physical locations of *D22S1*, *IGLV* (the immunoglobulin λ -chain variable region genes) and *SIS*, the gene for platelet-derived growth factor, are shown^{12,13,19}. The approximate location of BANF was inferred from the previous genetic mapping of *D22S1* relative to these two other markers¹².

in certain families. Because BANF is usually difficult to detect until symptoms appear, a predictive test could significantly modify patient management, prognosis and family planning. Individuals at risk found to be gene carriers could be carefully monitored for the earliest signs of tumour formation, facilitating successful surgical intervention before the development of the permanent neurologic dysfunction that can result from compression of critical nervous system structures.

The mapping of BANF to chromosome 22 by linkage analysis also provides the necessary foundation for isolating the gene defect by its location, an effort that should be facilitated by the availability of tumour tissue. The identification of interstitial deletions in acoustic neuromas or of structural alterations in BANF patients could serve to narrow down the range of possible positions of the BANF gene. Moreover, because the defective locus in BANF tumours is probably also the target of the somatic mutations causing sporadic acoustic neuromas, meningiomas and possibly some gliomas occurring in the general population, the cloning and characterization of the BANF gene could provide important insights into the pathogenesis of the most common tumours of the human nervous system.

We thank Dr R. White for providing the pMS3-18 probe and to the members of the BANF pedigree for their cooperation and assistance. Sample collection was partially supported by a contract from the National Neurofibromatosis Foundation. G.R. is supported in part by the Fonds de Recherche en Santé du Québec, and is a fellow of the MRC of Canada. J.L.H. is supported by an NIH training grant. B.R.S. is a fellow of the

National Neurofibromatosis Foundation. R.I.M. is recipient of an NINCDS Teacher-Investigator Career Development Award. J.F.G. is a Searle Scholar of the Chicago Community Trust. Funds for this work were provided by NIH grants, by the Julieanne Dorn Fund for Neurological Research and by the McKnight Foundation. Computing resources were provided by IUPUI computing services.

Received 20 May; accepted 20 July 1987.

1. Eldridge, R. *Neurology* **30**, 860-863 (1980).
2. Riccardi, V. M. *New Engl. J. Med.* **305**, 1617-1627 (1981).
3. Kanter, W. R., Eldridge, R., Fabricant, R., Allen, J. C. & Koerber, T. *Neurology* **30**, 851-859 (1979).
4. Seizinger, B. R. *et al. Science* **236**, 317-319 (1987).
5. Seizinger, B. R. *et al. Proc. natn. Acad. Sci. U.S.A.* (in the press).
6. Superneau, D. W., Wertelecki, W. & Conneally, P. M. *Am. J. hum. Genet.* **34**, 307 (1982).
7. Mulvihill, J. *Neurofibromatosis Res. Newsl.* **2**, No. 2 (National Neurofibromatosis Foundation, New York, 1986).
8. Barker, D., Schafer, M. & White, R. *Cell* **36**, 131-138 (1984).
9. Willard, H. F., Skolnick, M. H., Pearson, P. L. & Mandel, J. L. *Cytogenet. Cell Genet.* **40**, 360-489 (1985).
10. Ott, J. *Am. J. hum. Genet.* **28**, 528-529 (1976).
11. Hodge, S. F., Morton, L. A., Tideman, S., Kidd, K. K. & Spence, M. A. *Am. J. hum. Genet.* **31**, 761-762 (1979).
12. Julier, C. *et al. Cytogenet. Cell Genet.* **40**, 665 (1985).
13. Jhanwar, S. C., Neel, B. G., Hayward, W. S. & Chaganti, R. S. K. *Cytogenet. Cell Genet.* **38**, 73-75 (1984).
14. Knudson, A. G. *Proc. natn. Acad. Sci. U.S.A.* **68**, 820-823 (1971).
15. Cavenee, W. *Trends Genet.* **2**, 299-300 (1986).
16. Seizinger, B. R. *et al. Cell* **49**, 589-594 (1987).
17. Upadhyaya, M., Huson, S. & Harper, P. S. *Am. J. hum. Genet.* **39**, A171 (1986).
18. St George-Hyslop, P. *et al. Science* **235**, 885-889 (1987).
19. Tippet, P. & Kaplan, J.-C. *Cytogenet. Cell Genet.* **40**, 268-295 (1985).

Discrimination between antibodies to HIV and to related retroviruses using site-directed serology

Erling Norrby*, Gunnel Biberfeld†, Francesca Chiodi*, Agneta von Gegerfeldt*, Anders Naclér‡, Elliot Parks§ & Richard Lerner¶

* Department of Virology, Karolinska Institutet, School of Medicine, c/o SBL S-105 21 Stockholm, Sweden

† Department of Immunology, National Bacteriological Laboratory, 105 21 Stockholm, Sweden

‡ National Laboratory of Public Health, Bissau, Guinea Bissau

§ Johnson & Johnson Biotechnology Center, La Jolla, California 92038, USA

¶ Department of Molecular Virology, Research Institute of Scripps Clinic, La Jolla, California 92037, USA

Human immunodeficiency virus (HIV) strains can be separated into two types: HIV and HIV-related West African viruses. Site-directed serology using synthetic peptides offers possibilities for the determination of type-specific antibodies. A 22-amino-acid peptide with the sequence Ala-Ile-Glu-Lys-Tyr-Leu-Glu-Asp-Gln-Ala-Gln-Leu-Asn-Ala-Trp-Cys-Ala-Phe-Arg-Gln-Val-Cys representing a conserved region of the transmembranous protein of simian T-cell lymphotropic virus-type III (STLV-III; related to West African HIV) was used as antigen in an enzyme-linked immunosorbent assay (ELISA). In parallel, tests were performed with a pair of previously described peptides^{1,2}, including the homologous region of the glycoprotein (gp) 41 of the HIV strain HTLV-III_B. In tests with three groups of 20 sera it was shown that the different peptide ELISAs allowed a categorical distinction of antibodies to the two types of HIV. Tests using peptide antigens may provide excellent opportunities for large-scale testing for type-specific antibodies against HIV. The tests are simple, sensitive and specific and are readily standardized.

The discovery of the aetiology of acquired immune deficiency syndrome^{3,4} (AIDS) launched a uniquely rapid accumulation of knowledge about the disease and its causative agent. Details about the structure and genomic organization of HIV have been unravelled in a short time. Eventually this may lead to the emergence of means for effective therapeutic intervention and hopefully also immunoprophylaxis.

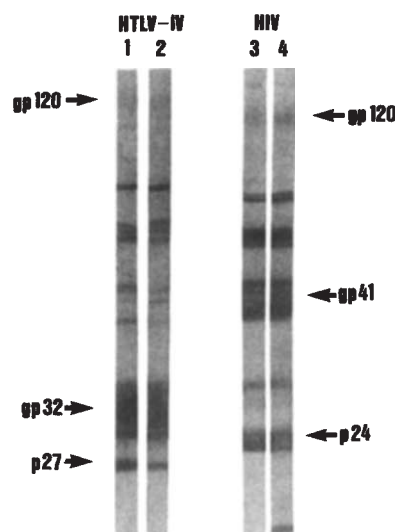
Possibilities for identification of HIV infections are gradually improving. The main emphasis in these endeavours concerns the development of specific, sensitive antibody assays. The tests used for screening are enzyme-linked immunosorbent assays^{5,6} (ELISA). Because these assays give a high frequency of non-specific reactions, confirmatory tests often have to be performed with ELISA-positive samples. These confirmatory tests include the Western blot technique which effectively identifies antibodies to the *gag* gene products p19 and p24 plus their precursor p55 and to the transmembranous protein gp41 (ref. 7) and the radioimmune precipitation assay (RIPA) which also allows a clear identification of the large glycoproteins, gp120 and its precursor gp160 (ref. 8).

One way of improving the identification of antibodies against



Fig. 1 Alignment of amino acids 586-620 of the transmembrane region of HIV (represented by the sequence of strain HTLV-III_B) and the corresponding region of STLV-III. The sequences included in the STLV-III peptides 1 and 2 and the previously described E32/E34 HIV peptides are indicated. The peptides were made by the solid-phase methods developed by Merrifield as described previously¹, using an automated peptide synthesizer (Applied Biosystems 430A). Peptide resins were cleaved by hydrogen fluoride, extracted and analysed for purity by high performance liquid chromatography (Waters Associates, Inc.) using a reverse phase Vydak C4 column.

Fig. 2 Western blot reactivity of two HTLV-IV antibody positive West African sera (numbers 1 and 2) with HTLV-IV antigen and of two HIV antibody positive Swedish sera (numbers 3 and 4) with HIV (HTLV-III_B) antigen. The HTLV-IV antigen was prepared from supernatants of HTLV-IV-infected HUT 78 cells by ultracentrifugation through 30% sucrose as described previously¹⁵. HTLV-IV pk 289 was obtained from Dr M. Essex, Harvard Medical School, Boston. The HIV antigen consisted of HTLV-III_B virions⁴ obtained as a gift from Dr R. C. Gallo. Western blot analysis was as described previously^{8,15}. Briefly, the viral proteins were separated by SDS-polyacrylamide gel electrophoresis (10 or 12% gel) and electrophoretically transferred onto nitrocellulose. Nitrocellulose strips were incubated with human serum diluted 1:100 followed by biotinylated sheep anti-human immunoglobulin, streptavidin and biotinylated horseradish peroxidase (Amersham). The colour was developed with *o*-dianisidine.



HIV is to use site-directed serology employing synthetic peptides⁹. In the search for immunodominating linear antigenic sites in HIV proteins several groups have focused on a small part of the N-terminal end of the gp41 protein^{1,2,10,11}. Various peptides representing the exposed, conserved region of amino acids 586–620 have been found to be highly effective antigens in ELISAs. Optimal results in terms of specificity and sensitivity were obtained with two partly overlapping peptides representing amino acids 589–610 (refs. 1, 2, 12). This test, developed at the Johnson & Johnson Biotechnology Center is referred to as the E32/E34 peptide ELISA.

Serodiagnosis of HIV infection has become complicated by the discovery of HIV-related viruses with markedly distinct envelope glycoproteins. Thus there is a separate group of HIV-related viruses including the strains HTLV-IV¹³, LAV-2¹⁴ and SBL-6669¹⁵, all having antigenically closely-related glycoproteins. All these strains have been isolated from individuals of West African origin. In tests with a small number of sera containing antibodies to the West African HIV-related viruses it was found that no reaction could be detected in the E32/E34 peptide ELISA¹². The immunodominant antigenic site in the region of gp41 covered by these two peptides thus has different immunogenic properties in HIV and HIV-related viruses.

Attempts were therefore made to develop a peptide ELISA allowing identification of antibodies against the group of HIV-related viruses. We assumed two things: first, that the immunodominant region in the gp41 protein of HIV would have its counterpart in HIV-related viruses; second, that the cross-reaction between the different HIV-related viruses infecting man would include the simian virus STLV-III, shown in one study to be closely related to HTLV-IV¹³. Recent data indicate that the STLV-III genome has characteristics essentially indistinguishable from those of the HTLV-IV genome¹⁶.

The genomic nucleotide sequence of STLV-III has recently been determined¹⁷. A comparison of the amino acid sequence of the N-terminal part of the gp41 protein of HIV and the corresponding region of STLV-III was performed. With some lateral adjustments including the parallel positioning of two cysteine residues, an alignment which gave identity in 18 of 32 amino acid positions was found (Fig. 1). Two overlapping peptides (1 and 2) covering the identified stretch of 32 amino acids in STLV-III (Fig. 1) were made at Johnson & Johnson Biotechnology Center using the Merrifield solid-phase method as described previously¹.

Preliminary tests with two sera containing high titres of antibodies against HIV-related West African viruses were performed with each peptide separately and as a mixture. Peptide 1 (Fig. 1) gave high extinction values with both sera, whereas lower

values were obtained with peptide 2. When the peptides were mixed, reduced extinction values compared to those obtained with only peptide 1 were seen. It was therefore decided to use only peptide 1 in the ELISA.

Three groups of sera were used to evaluate the reactivity in ELISA of the STLV-III peptide 1 and the previously studied HIV E32/E34 peptides^{1,2,12}. The first group consisted of 20 sera from HTLV-IV antibody-positive West African subjects, 18 of them attending an outward clinic in Bissau, Guinea-Bissau, for examination of suspected tuberculosis and the other two being healthy immigrants to Sweden (reported in ref. 18 as cases B and C). All of the sera reacted with HTLV-IV by ELISA using disrupted virions as antigen (G. Biberfeld *et al.*, unpublished data) and with the presumed transmembrane protein gp32 of HTLV-IV in Western blot analysis¹⁵ (Fig. 2), but not with gp41 of HIV. The second group included sera from 20 HIV antibody-positive asymptomatic Swedish subjects, reacting in the Western blot assay with gp41 and other antigens of HIV (Fig. 2) but not with gp32 of HTLV-IV. The third group consisted of 20 sera from Swedish blood donors negative for HIV and HTLV-IV antibodies by ELISA (Organon and Wellcome anti-HIV ELISAs and HTLV-IV virion ELISA). Figure 3 gives the results of the peptide ELISA with different sera.

The 20 sera containing antibodies to West African HIV-related viruses gave pronounced reactions with the STLV-III peptides, but, in agreement with previous observations¹², no reaction was seen with the HIV-specific E32/E34 peptides except in one case. The reaction with the latter peptide in this case gave an E_{492} value of 0.2 compared with the homologous value of 1.5. In contrast, the 20 sera with antibodies against HIV all reacted exclusively with the E32/E34 peptides, giving extinction values of <0.15 with STLV-III peptides, with a single exception. The serum which showed a cross-reaction gave an E_{492} value of 1.5 with the E32/E34 peptides compared with a value in the heterologous test of 0.45. The 20 blood-donor sera did not react with any of the peptides.

The requirements on serological tests for identification of antibodies against human lentiviruses are that they should be specific and sensitive. It is also important that they allow distinction between the hitherto-established two types of these viruses. The STLV-III peptide ELISA test presented in this report and the previously described E32/E34 peptide ELISA for HIV^{1,2,12} appear to satisfy these requirements well.

The N-terminal region of the transmembraneous protein represented by the peptides examined is apparently a highly immunodominant site in this protein both in HIV and in STLV-III and related viruses. The presence of cysteines in the peptides may contribute to the formation of a configuration mimicking

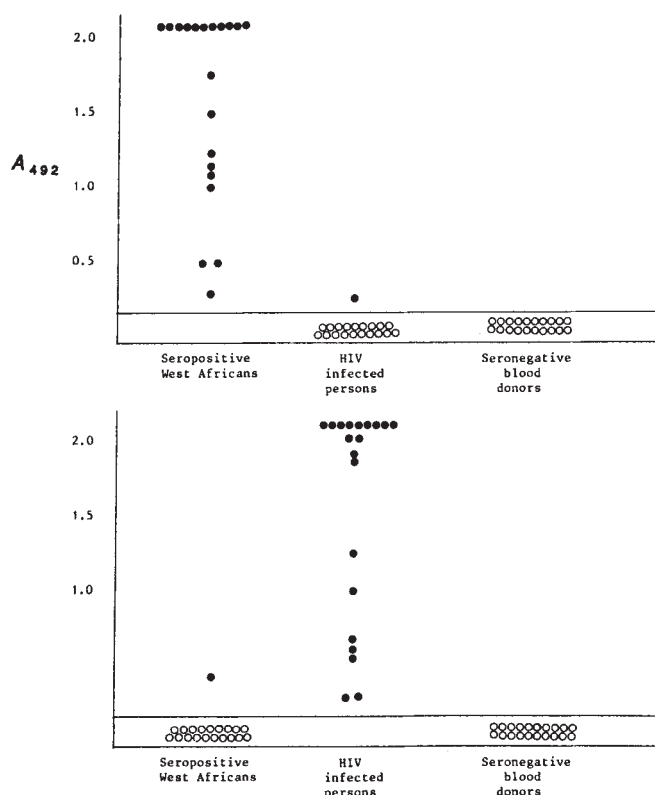


Fig. 3 Antibody responses in a group of 20 sera from individuals in West Africa with antibodies to HIV-related viruses, 20 sera from HIV-infected individuals and 20 sera from blood donors without antibodies to HIV. Top, STLTV-III peptide ELISA; bottom, HIV peptide ELISA. Closed symbols denote positive samples; open symbols negative samples. The peptide ELISAs were performed as previously described^{1,12}. Flat-bottomed micro-ELISA plates (Dynatech Immunolon II) coated with 200 μ l of 5 μ g ml⁻¹ of each peptide in 0.1 M sodium bicarbonate buffer, pH 9.6, were used. After coating, the plates were blocked for 90 min with 10% BSA to reduce non-specific binding to the solid phase and dried, and human serum diluted 1:10 was added to the wells. After incubation for 30 min at 37 °C horseradish peroxidase-conjugated mouse monoclonal antibody to human IgG (Ortho Diagnostic System, Inc.) diluted 1:3,500 was added to the plates. The plates were incubated for another 30 min at 37 °C, a solution containing the substrate *o*-phenylenediamine added and the plates read at 492 nm. The colour reaction was stopped by addition of 4-N H₂SO₄. Sera were considered positive when the optical density values were above 0.150 (the mean optical density of negative samples plus six standard deviations).

that of the native protein. The test plates with the attached peptides have a very high density of antigenic sites. In spite of this the background extinction values are remarkably low, which makes it possible to examine sera at low dilution. This probably also applied to 'sticky' sera from African although we have not

yet shown that the signal to noise ratio is also high in these cases. Taken together these considerations highlight the sensitivity of the peptide ELISA.

Peptide ELISA tests are antigenic-site specific. It is this that allows the STLTV-III peptide ELISA and the previously presented E32/E34 ELISA to discriminate categorically between antibody responses to HIV-related viruses and HIV. As concerns the representation of target viral protein the *env* product is most frequently detected by antibodies from exposed individuals. There is evidence that antibodies to envelope antigens appear before antibodies to internal components in the primary infection¹⁹. Further, those patients who lose antibodies in advanced stages of AIDS lose antibody to internal but not envelope virion components²⁰. The accumulated experience of use of the E32/E34 ELISA for identification of antibodies to HIV indicates that the test is superior to other available ELISAs¹. If further testing of the STLTV-III peptide confirms that it can match the E32/E34 ELISA at specific antibody detection, it will be possible to develop a simple human lentivirus serological test with dual specificity. The availability of this kind of test will simplify screening of large groups of individuals, for example, blood donors.

The situation regarding human lentivirus infections in Africa appears catastrophic²¹. There is an emergency need for introduction of simplified, inexpensive, sensitive and specific tests for simultaneous but separate identification of antibodies to HIV and HIV-related viruses. Peptides used in site-directed ELISA can be chemically synthesized in gram quantities under highly standardized conditions and provide a source of antigen of reproducible integrity. Plates containing peptides are highly stable. It would seem therefore that peptide ELISAs are excellently suited for use in developing countries. It may now be possible to characterize the extent of epidemics of HIV and HIV-related West African retroviruses, to identify the occurrence of double infections with these subgroups of virus and to define the association of the latter viruses with disease development, a question still under debate²².

After submission of this report it became apparent that there was an error in the STLTV-III sequence used for production of peptides. Thus, in the first position left open in the sequence (Fig. 1) there is in fact a Gly homologous to that in the HIV sequence. An STLTV-III peptide containing this missing Gly has been tested in ELISA with a few sera containing antibodies against HIV-related West African viruses. This peptide appears to perform slightly better than the peptide used here. Of further interest is the fact that the nucleotide sequence of LAV-2 was published recently²⁴. A comparison of the homologous sequences in the selected part of the transmembranous protein of STLTV-III and LAV-2 revealed identity in as many as 28 of 33 amino acids, emphasizing the conservation of this antigenic site in HIV-related West African viruses.

We thank Drs R. Gallo and G. Franchini for exchange of scientific information before publication and Mrs Susanne Gröndahl and Alice Whalley for excellent technical assistance. Grants to support this work were obtained from the Swedish Medical Research Council.

Received 5 May; accepted 22 July 1987.

- Smith, R. S. *et al. J. clin. Microbiol.* **25**, 1498-1504 (1987).
- Rosen J. *et al. in Modern Approaches to Vaccines* (eds Chanock, R., Ginsberg, H., Lerner, R. & Brown, F.) (Cold Spring Harbor Laboratory Press, New York).
- Barré-Sinoussi, F. *et al. Science* **220**, 868-871 (1983).
- Popovic, M., Sarngadharan, M. G., Read, E. & Gallo, R. C. *Science* **224**, 497-500 (1984).
- Abb, J. *AIDS Res.* **2**, 93-97 (1986).
- Resnick, H. W. *et al. Lancet* **ii**, 483-486 (1986).
- Sarngadharan, M. G. *et al. Science* **224**, 506-508 (1984).
- Chiodi, F. *et al. AIDS Res. hum. Retroviruses* **3**, 165-176 (1987).
- Lerner, R. A. *Adv. Immunol.* **36**, 1-44 (1984).
- Wang, J. J. G., Steel, S., Wisniewski, R. & Wang C.Y. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6159-6163 (1986).
- Gnann, J. W., Schwimbeck, P. L., Nelson, J. A., Truax, B. & Oldstone, M. B. A. *J. inf. Dis.* **156**, 261-267 (1987).

- Chiodi, F. *et al. J. med. Virol.* **23** (in the press).
- Kanki, P. J. *et al. Science* **232**, 238-243 (1986).
- Clavel, F. *et al. Science* **233**, 343-346 (1986).
- Albert, J. *et al. AIDS Res. hum. Retroviruses* **3**, 3-10 (1987).
- Kornfeld, H., Riedel, N., Viglianti, G. A., Hirsch, V. & Mullins J. J. *Nature* **326**, 610-613 (1987).
- Franchini, G. *et al. Nature* **328**, 539-543 (1987).
- Biberfeld, G. *et al. Lancet* **ii**, 1330-1331 (1986).
- Gaines, H. *et al. Lancet* **i**, 1249-1253 (1987).
- Montagnier, L. *Virology* **144**, 283-289 (1985).
- Quinn, T. C., Mann, J. M., Curran, J. W. & Piot, P. *Science* **234**, 955-961 (1986).
- Kanki, P. J. *et al. Science* **236**, 827-831 (1987).
- Ratner, L. *et al. Nature* **313**, 277-283 (1985).
- Guyader, M. *et al. Nature* **326**, 662-669 (1987).

A novel population of T-cell receptor $\alpha\beta$ -bearing thymocytes which predominantly expresses a single V_β gene family

B. J. Fowlkes*, A. M. Kruisbeek§, H. Ton-That*, M. A. Weston§, J. E. Coligan†, R. H. Schwartz‡ & D. M. Pardoll‡

* Laboratory of Microbial Immunity, † Biological Resources Branch, and ‡ Laboratory of Cellular and Molecular Immunology, National Institute of Allergy and Infectious Diseases; § Biological Response Modifiers Program, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Recent studies have demonstrated that CD3 is expressed on a subset of thymocytes with a $CD4^-CD8^-$ (double negative) phenotype^{1,2}. At least some of these cells bear the CD3-associated $\gamma\delta$ T-cell receptor (TCR $\gamma\delta$)^{3,4}. Here we describe a second subset of double negative thymocytes which expresses CD3-associated $\alpha\beta$ receptors (TCR $\alpha\beta$). Surprisingly, these cells express predominantly the products of a single V_β gene family ($V_\beta 8$). These $CD4^-CD8^-$, TCR $\alpha\beta^+$ cells appear relatively late in ontogeny (between birth and day 5 of life) and thus are unlikely to be the precursors to the TCR $\alpha\beta$ -bearing cells ($CD4^+CD8^-$ and $CD4^-CD8^+$) already present at birth. They can be selectively expanded *in vitro* by stimulation with a monoclonal antibody to $V_\beta 8$ (F23.1)⁵ in the presence of interleukin 1 (IL-1). We propose that this cell type is a unique T-cell population distinguishable from typical TCR $\alpha\beta^+$ T cells by its $CD4^-CD8^-$ phenotype and a restricted TCR V_β repertoire. Analysis of the unique phenotype of these cells suggests that they may represent the normal counterpart of the defective $CD4^-CD8^-$ T cells found in the *lpr* autoimmune mouse.

In previous studies, we have shown that a subpopulation of adult murine thymocytes (3–5%) isolated by cytotoxic elimination with anti-CD4, CD8 and CD5 (Ly-1) antibodies contains the precursors to all the major subpopulations of thymocytes⁶. This subpopulation was shown to express TCR β and γ but no TCR α messenger RNA transcripts^{7–9}. Immunoprecipitation of detergent lysates derived from these surface labelled cells with a monoclonal anti-CD3 antibody (145-2C11)^{2,10}, or an anti-TCR γ antiserum identified a CD3-associated heterodimer containing the γ chain, disulphide-linked to a second protein, termed TCR δ^3 . In the present study, we compared the surface expression of TCR on dull-CD5, $CD4^-CD8^-$ with that on total $CD4^-CD8^-$ thymocytes which, additionally, contain a population with a bright CD5, $CD4^-CD8^-$ phenotype. Immunoprecipitations of ¹²⁵I-labelled cell lysates of these two subpopulations run on SDS-polyacrylamide gel electrophoresis (SDS-PAGE) under reducing conditions are shown in Fig. 1. Precipitation with anti-CD3 reveals that, in addition to the γ and δ proteins (relative molecular masses 35,000 (M_r 35K) 35 and 45K, respectively) found previously³ for dull-CD5 $CD4^-CD8^-$ thymocytes (Fig. 1a), significant amounts of 37–43K protein, suggestive of an $\alpha\beta$ heterodimer, are precipitated from labelled lysates of total double-negative thymocytes (Fig. 1b). The identity of these proteins is confirmed by their ability to be immunoprecipitated with anti- C_α and anti- C_β , but not anti- C_γ peptide serum (Fig. 1b). Immunoprecipitation of the 37–43 K material with these antisera is inhibited in the presence of the corresponding C_β and C_α peptides used to raise the antisera. These results indicate that a subset of the bright CD5 double-negative thymocytes bear a CD3-associated TCR $\alpha\beta$ heterodimer.

To analyse the distribution of CD3-associated TCR at the single cell level, isolated $CD4^-CD8^-$ thymocytes were analysed for coordinate expression of CD3, CD5 and TCR $V_\beta 8$ (refs 5,

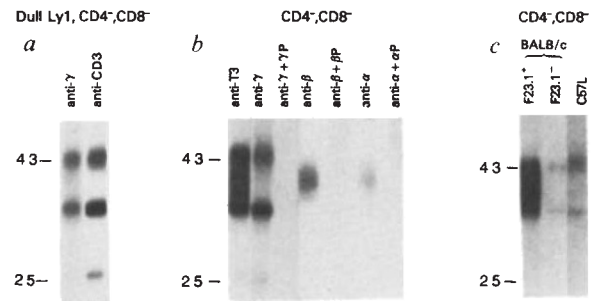


Fig. 1 Differential expression of two T-cell receptors in $CD4^-CD8^-$ thymocytes. Fractionated thymocytes were radioiodinated and lysed in digitonin or Nonidet-P-40 (NP-40). Lysates of: *a*, B6 dull-CD5, $CD4^-CD8^-$; *b*, total B6 $CD4^-CD8^-$; *c*, total BALB/c $CD4^-CD8^-$; BALB/c $CD4^-CD8^-$, F23.1⁺; and C57L $CD4^-CD8^-$ thymocytes were immunoprecipitated with anti-CD3, anti- γ -chain, anti- α -chain, or anti- β -chain antibodies in the presence or absence of the relevant peptide as designated above each lane. All samples were run on SDS-PAGE gels under reducing conditions. Size markers, M_r , in thousands.

Methods. $CD4^-CD8^-$ cells were isolated by sequential cytotoxic elimination using rat monoclonal antibodies to CD4 (LICR-LAU, RL172.3)²⁷ and mouse monoclonal antibodies to CD8.2 (19/178)²⁸, plus a 1:12 dilution of Low-Tox-M rabbit complement (Cedar Lane, Ontario, Canada). Isolation of dull-CD5 $CD4^-CD8^-$ thymocytes involved an additional cytotoxic elimination with mouse antibodies to CD5.2 (Ly-1.2) (New England Nuclear)⁶. F23.1⁺ $CD4^-CD8^-$ thymocytes from BALB/c mice were obtained by binding F23.1 antibody to isolated $CD4^-CD8^-$ thymocytes at 4 °C for 30 min. After two washes with medium, the cells were incubated with affinity purified rabbit anti-mouse immunoglobulin (Zymed) and Low-Tox-M rabbit complement (1:12) at 37 °C for 30 min. Viable cells were collected after centrifugation through Ficoll-Paque (Pharmacia, Piscataway, NJ). $CD4^-CD8^-$ thymocytes used as a positive control were obtained in the same manner, using isotype-matched mouse mAb to Ly-5.2 (New England Nuclear). Purity was >98% by FC analysis using appropriate FITC-labelled isotype- and species-specific second antibodies. Cells ($1-5 \times 10^7$) were ¹²⁵I-labelled using lactoperoxidase as previously described³. Cells were lysed in 0.5% NP-40 (for anti- α -chain, β -chain and γ -chain precipitations) or 1% digitonin buffer (for anti-CD3 immunoprecipitations) containing phenylmethylsulphonyl fluoride, leupeptin, pepstatin and chymostatin. Samples were pre-cleared with rabbit anti-chicken immunoglobulin. Anti-CD3 immunoprecipitations were performed with culture supernatants from the 145-2C11 hybridoma¹⁰, rabbit anti- α peptide (residues 113–124) serum, rabbit anti- β peptide (C-terminal C β 2 octapeptide) serum, and anti- γ peptide serum. Rabbit anti- γ -chain serum was affinity purified on a γ -peptide column as described⁴. Immune complexes were collected with protein A agarose beads and run under reducing conditions on 10% SDS-PAGE gels. Gels were dried and labelled, and proteins were visualized using Kodak XAR-5 autoradiographic film at –70 °C with enhancing screens.

11–14) using two-colour fluorescence flow cytometry (FC). The TCR $V_\beta 8$ -specific monoclonal antibody used, F23.1, reacts with all three members of the $V_\beta 8$ family¹⁵ and stains 18–25% of peripheral T cells in various mouse strains possessing this gene segment⁵. When $CD4^-CD8^-$ thymocytes from BALB/c mice were analysed, ~10% of these cells are F23.1⁺, ~20% CD3⁺ and ~16% are CD5-bright. Two-colour FC reveals that all $V_\beta 8^+$ double negative thymocytes coordinately express CD3 (Fig. 2a) and high levels of CD5 (Fig. 2b). Indeed, the $V_\beta 8^+$ cells co-expressed high levels of CD5 in all strains analysed. These results from FC analyses are consistent with the absence of TCR $\alpha\beta$ revealed by biochemical analysis of dull CD5, $CD4^-CD8^-$ thymocytes after CD5-bright cells are eliminated (Fig. 1a, b). Figure 2c demonstrates further that virtually all CD5-bright, double-negative thymocytes are CD3⁺. As we show here (Fig. 1a) and as we described previously³, the few CD3⁺ cells (~4%) which express lower levels of CD5 bear only TCR $\gamma\delta$.

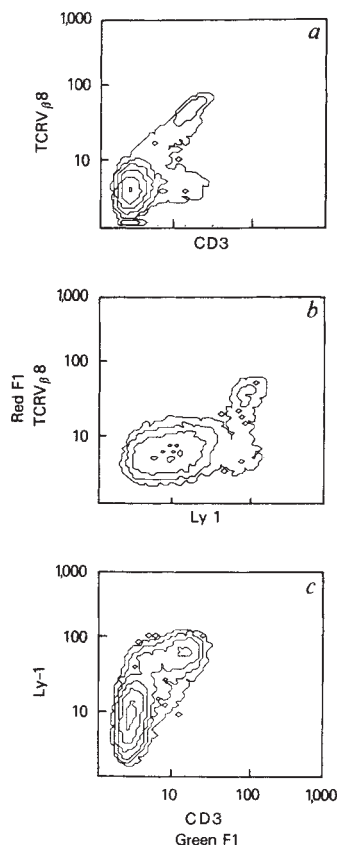


Fig. 2 Two-colour fluorescence FC analysis for coordinate expression of $V_{\beta}8$, CD3 and CD5 by $CD4^{-}CD8^{-}$ adult thymocytes. Double-negative thymocytes from BALB/c mice were isolated as described in Fig. 1 and stained for *a*, CD3 and $V_{\beta}8$, *b*, CD5 and $V_{\beta}8$ and *c*, CD3 and CD5. Percent positive $CD4^{-}CD8^{-}$ thymocytes from BALB/c: for CD3, 20.1%; $V_{\beta}8$, 10.5%; CD5-bright, 17.4%; from B6: CD3, 7.6%; $V_{\beta}8$, 2.6%; CD5-bright, 4.8%; from B10: CD3, 11.9%; $V_{\beta}8$, 4.4%; CD5-bright, 7.5%; from C3H: CD3, 10.7%; $V_{\beta}8$, 3.4%; CD5-bright, 5.9%; from B10.D2: CD3, 9.3%; $V_{\beta}8$, 4.7%; CD5-bright, 8.0%; and from C57L: CD3, 9.4%; $V_{\beta}8$, 0.0%; CD5-bright, 6.5%. Percent $V_{\beta}8^{+}$ splenic T cells for BALB/c, 29.8%; B6, 20.6%; B10, 21.4%; C3H, 28.5%; B10.D2, 21.3%; C57L, 0.0%. Mice were 5–8 weeks old. Data are representative of three separate experiments.

Methods. Immunofluorescence staining and washing was carried out as previously described⁶. Thymocytes were stained in sequential steps. Reagents used included the hamster mAb to CD3, 145-2C11 (ref. 10), followed by fluorescein isothiocyanate (FITC)-labelled mouse antibody to hamster immunoglobulin, biotin-conjugated mouse monoclonal anti-TCR $V_{\beta}8$ antibody, F23.1⁵ followed by Texas red-labelled avidin (Jackson ImmunoResearch), rat monoclonal anti-mouse CD5 antibody, 53.7.3 (ref. 29) followed by FITC-labelled species-specific goat anti-rat immunoglobulin (Kirkegaard & Perry), and biotin-conjugated rat monoclonal anti-mouse CD5 antibody (Becton-Dickinson) followed by Texas red-labelled avidin. Background values were obtained using isotype-matched, irrelevant antibodies, and have been subtracted from the reported values. FC analysis was performed using a B-D Dual Laser FACS II interfaced to a Digital Equipment Corporation PDP 11/24 computer as previously described^{2,6,30}. Viable cells (10^5) were analysed by forward light scatter and propidium iodide gating.

Taken together, these analyses reveal that at least one-half of the $CD3^{+}$ double-negative thymocytes in BALB/c mice use V regions encoded by the $V_{\beta}8$ gene family. If the $CD3^{+}$ cells in the CD5-dull, double negatives are subtracted from the values (as they bear only $\gamma\delta$ receptors), the proportion of $CD3^{+}$ cells which use $V_{\beta}8$ is even higher (~63%). Moreover, if some of the CD5 bright cells have their CD3 associated with $\gamma\delta$ receptors,

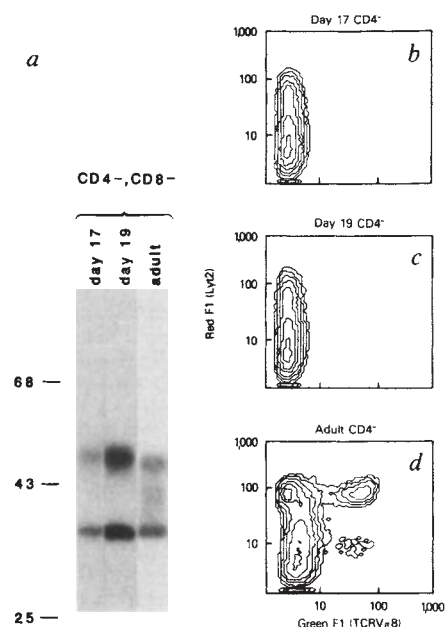


Fig. 3 Failure of C57BL/6 $CD4^{-}CD8^{-}$ fetal thymocytes to express TCR $\alpha\beta$. *a*, $CD4^{-}CD8^{-}$ day 17 and day 19 fetal or adult thymocytes were radioiodinated, lysed in digitonin, and immunoprecipitated with anti-CD3 as described in Fig. 1. Two-colour FC analysis of $CD4^{-}$ fraction from *b*, day 17 fetal; *c*, day 19 fetal; and *d*, adult thymocytes for coordinate expression of CD8 and TCR $V_{\beta}8$ were performed as described in Fig. 2.

Methods. Fetal thymocytes were obtained from timed pregnant B6 mice. $CD4^{-}CD8^{-}$ or $CD4^{-}$ fetal thymocytes were isolated by cytotoxic elimination as described in Fig. 1.

then the proportion of $\alpha\beta^{+}$ double negative cells that use $V_{\beta}8$ could be even higher.

These results raised the possibility that virtually all the TCR $\alpha\beta^{+}$ CD5-bright, double-negative thymocytes use a β chain derived from the $V_{\beta}8$ gene family. Because no monoclonal pan anti-murine $\alpha\beta$ antibody is available, we addressed this question by depleting $CD4^{-}CD8^{-}$ thymocytes from BALB/c mice of $V_{\beta}8^{+}$ cells by lysis using F23.1 antibody. Immunoprecipitations with anti-CD3 revealed that virtually all the 37–43 K material corresponding to the $\alpha\beta$ heterodimer is eliminated by the F23.1 antibody when compared with double-negative thymocytes treated with irrelevant antibodies (Fig. 1c). As expected, the 35 K and 45 K proteins (TCR γ and δ) are unaffected. These results indicate that the bulk of the TCR $\alpha\beta$ receptors on double-negative thymocytes in BALB/c mice contain β chains encoded by a single V_{β} gene family: $V_{\beta}8$, and that the $CD3^{+}$, CD5-bright cells remaining after F23.1 antibody and complement lysis bear $\gamma\delta$ receptors. Indeed, immunoprecipitations using anti-CD3 of lysates from the CD5-bright, double-negative thymocyte subset isolated by electronic cell sorting reveal both TCR $\alpha\beta$ and $\gamma\delta$ (data not shown).

As is true for peripheral T cells, the $V_{\beta}8$ expression by double-negative thymocytes varies among different strains of mice (Fig. 2 legend). Nevertheless, the proportion of $CD3^{+}CD5$ -bright, double-negative cells expressing $V_{\beta}8$ gene products (54–60%) is much higher than has been reported for peripheral T cells (<30%)⁵. In confirmation that the F23.1 antibody binding is due to $V_{\beta}8$ expression rather than a cross-reaction, FC analysis of double-negative thymocytes from strains lacking the $V_{\beta}8$ gene family (C57L, SJL) show no F23.1 staining (Fig. 2, legend). Interestingly, anti-CD3 immunoprecipitation of lysates from surface-iodinated, double-negative thymocytes of C57L mice reveals virtually no 37–43K material (Fig. 1c); similarly, no 37–43K material is detectable in SJL lysates (data not shown).

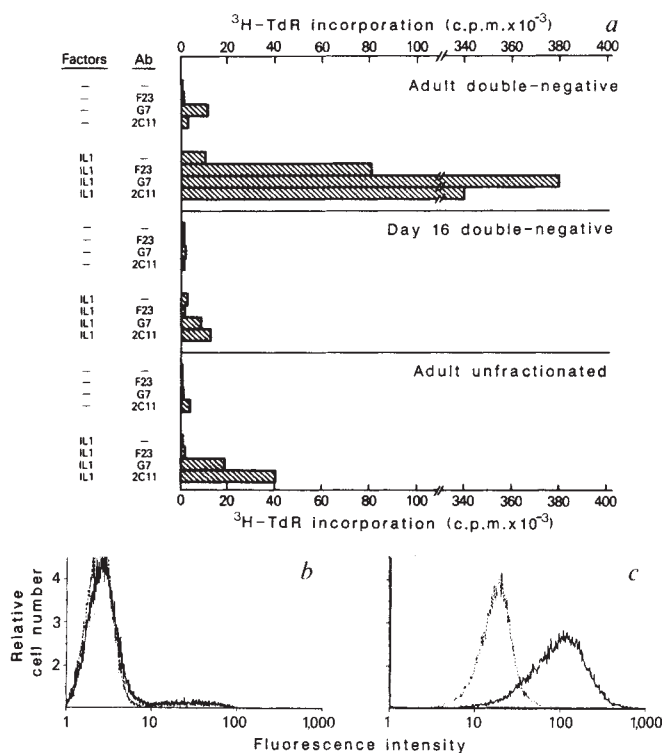


Fig. 4 Stimulation and selective growth of $\text{CD4}^+\text{CD8}^-$ adult thymocytes using monoclonal antibodies to TCR $\text{V}_\beta 8$ (F23.1) plus IL-1. **a**, Unfractionated B6 $\text{CD4}^+\text{CD8}^-$ adult, or day-16 fetal thymocytes were cultured in the presence of antibody to mouse TCR $\text{V}_\beta 8$, Thy-1.2 (G7)¹⁶, or CD3 (145-2C11)¹⁰ in the presence or absence of IL-1. Proliferation was determined by incorporation of ^3H -TdR. Data is representative of three separate experiments. **b**, Single-colour fluorescence flow cytometry of freshly isolated $\text{CD4}^+\text{CD8}^-$ adult thymocytes. **c**, $\text{CD4}^+\text{CD8}^-$ adult thymocytes cultured for 3 days with F23.1 antibody plus IL-1, followed by a 2-day rest in the absence of antibody. Cells were stained with FITC-labelled antibody F23.1 (—) compared with control FITC-labelled, isotype-matched, irrelevant antibody, Ly-5.2 (New England Nuclear) (----).

Methods. $\text{CD4}^+\text{CD8}^-$ thymocytes were isolated as described in Fig. 1. Cells were cultured at 3×10^5 per well in 96-well round-bottom plates as previously described³¹ using antibodies F23.1 at $5 \mu\text{g ml}^{-1}$, G7 at $5 \mu\text{g ml}^{-1}$, or a diluted supernatant of 145-2C11. Recombinant IL-1 and IL-2 (a gift of Hoffman-LaRoche) were at concentrations of 20 U ml^{-1} and 50 U ml^{-1} , respectively. Cells were cultured for 3 days; $1 \mu\text{Ci } ^3\text{H}$ -TdR per well was added 18 h before harvesting. Immunofluorescence staining and FC analysis was carried out as described in Fig. 2.

In earlier studies, we demonstrated by *in vivo* transfer experiments that double-negative thymocytes contain the precursors to the mature thymocytes ($\text{CD4}^+\text{CD8}^-$ and $\text{CD4}^+\text{CD8}^+$)⁶. The finding of TCR $\alpha\beta$ within double-negative thymocytes raises the possibility that these receptor-bearing cells may serve as precursors to single-positive $\alpha\beta^+$ thymocytes. Because the early stages of T-cell development are more prominent and maturation more synchronized in the fetal thymus, we reasoned that, if these $\alpha\beta$ -bearing double-negative cells are precursors, they should be more prevalent in the fetal thymus. On the contrary, Fig. 3 shows that $\alpha\beta$ -bearing double-negative thymocytes are undetectable in the fetus by either biochemical or FC analysis. We find no TCR $\alpha\beta$ by anti-CD3 immunoprecipitation of C57BL/6 (B6) day 17 or day 19 fetal thymocytes after depletion of the CD4^+ and CD8^+ cells (Fig. 3a). Likewise, FC analysis reveals no F23.1⁺ cells in fetal double-negative thymocytes (Fig. 3b-d). Even analysis of fetal thymocytes from BALB/c, which contain more adult F23.1⁺ double-negative cells (10-

11%) than B6 (2-3%) reveals no F23.1⁺ cells before birth (data not shown). On the contrary, F23.1⁺ double-negative thymocytes are first detected at very low levels in BALB/c mice on the day of birth and at day 4-5 of life in B6 mice (~0.5%). Adult levels of these cells are found in B6 mice by 2 weeks of life. Nevertheless, we believe these double-negative, $\alpha\beta^+$ cells are most probably thymically derived as they appear late in fetal thymic organ cultures as determined by FC analysis (S. Sharrow and D. DeLuca, personal communications). The late appearance of these cells suggests that they are not the precursors at least of those CD4^+ or CD8^+ functional cells that appear before birth. Transfer studies are underway to rule out the possibility that the $\alpha\beta^+$ double-negative cells are the precursors to any later-appearing single-positive thymocytes.

Activation of mature, functional T cells bearing TCR $\alpha\beta$ can be achieved not only by antigen but also by antibodies to TCR, to associated CD3 components, or to the Thy-1 antigen. To determine whether double-negative $\alpha\beta^+$ thymocytes express a functional receptor, we tested the ability of these cells to be stimulated using F23.1 antibody with various combinations of the lymphokines, interleukin 1 (IL-1), interleukin 2 (IL-2), or irradiated spleen cells. In all cases, the highest proliferation of adult double-negative thymocytes was obtained with antibody in the presence of IL-1 (data not shown). Figure 4a shows that double negative adult thymocytes proliferate also with G7 (anti-Thy-1.2)¹⁶ and anti-CD3 antibodies when used in combination with IL-1. Because both $\alpha\beta^-$ and $\gamma\delta$ -receptor-bearing cells are contained in this subpopulation, it is impossible to say which cells are stimulated using these culture conditions. Any subset of cells may be expanding as a result of cell interactions or lymphokine production. But in the case of F23.1 stimulation, the F23.1⁺ cells appear to be the main responding population, as essentially all of the cells are F23.1⁺ and remain double-negative by FC analysis after 3 days of culture (Fig. 4b,c). The failure of day-16 fetal thymocytes to respond to F23.1 antibody in these proliferation assays is consistent with the FC analyses indicating no detectable F23.1⁺ cells at this stage of development (Fig. 3a). These same day 16 double-negative fetal thymocytes proliferate well in response to anti-CD3 in the presence of IL-2 and irradiated spleen cells, demonstrating that a subset of these fetal cells can proliferate (by virtue of their $\gamma\delta$ receptors) using different culture conditions (data not shown).

The ability to stimulate $\text{V}_\beta 8^+$, double-negative thymocytes and selectively grow these cells in culture with F23.1 antibody suggests that these $\alpha\beta^+$ cells, like their CD4^+ or CD8^+ single-positive counterparts, can be stimulated to proliferate through perturbation of the TCR. The fact that these $\alpha\beta^+$ double-negative thymocytes express a functional T-cell receptor, as well as their late appearance in ontogeny, suggests that these cells may be fully differentiated. Moreover, the virtual absence of $\alpha\beta^+$ double-negative thymocytes in C57L argues against a precursor role for these cells. On the contrary, these cells may represent a distinct lineage, separate from the bulk of $\alpha\beta$ -bearing CD4^+ or CD8^+ cells. Further phenotypic analysis of this $\text{V}_\beta 8^+$ $\text{CD3}^+\text{CD4}^-\text{CD8}^-$ subpopulation indicates that these cells are Thy-1⁺, Bgp⁺^{17,18}, CD5^+ , PC.1^- ¹⁹, HSA^- ²⁰, Ly6c^- ²¹, IL-2R^- ²² and B220^- ²³. Although the function of most of these surface antigens is not known, this phenotypic characterization further distinguishes these cells from typical CD4^+ or CD8^+ T cells. Strikingly, this phenotype is similar to that of the defective $\text{CD4}^+\text{CD8}^-$ $\alpha\beta^+$ cells found in the lymph nodes of *lpr* autoimmune mice^{24,25}. Mice homozygous for the mutant *lpr* gene develop multiple autoantibodies, glomerulonephritis, vasculitis, and massive lymphadenopathy. The predominant cell type in the lymph nodes from these mice express surface antigens usually restricted to the T cell but also express some B-cell markers²⁵. The B-cell markers could be aberrantly expressed on *lpr* cells as there is no rearrangement of immunoglobulin heavy-chain genes. Of note, the majority of cells within the abnormal *lpr* lymph node population expresses products of the $\text{V}_\beta 8$ gene family. Although

the CD4⁺CD8⁺ $\alpha\beta^+$ cells described here do not express B-cell surface antigens, a fraction of these cells becomes positive for the B-cell markers expressed on *lpr* cells, B220 and PC.1, upon activation with F23.1 antibody. It is, therefore, possible that the $\alpha\beta^+$, double-negative thymocytes described here could be the source of the defective cells found in *lpr* mice. In fact, thymectomy of *lpr* mice prevents the appearance of the abnormal cells²⁶. Our preliminary studies indicate the presence of V β 8⁺ CD4⁺CD8⁺ cells in activated lymph nodes of normal mice. The relationship between these activated peripheral lymph node cells and V β 8⁺ CD8⁺CD4⁺ thymocytes from normal mice is under investigation.

The possibilities remain that $\alpha\beta^+$, double-negative thymocytes serve as precursors, function in some regulatory capacity intrathymically, or may even represent some by-product of aberrant differentiation. Clearly, the most striking characteristic of this novel T-cell population is its highly restricted repertoire of V β gene usage such that a single V β gene family (V β 8) is expressed on the majority of the cells. Thus, the receptor repertoire of these cells is apparently subject to some sort of stringent selective pressure. Certainly the existence of this unique thymocyte population warrants further investigation of the mechanism responsible for the restricted repertoire and the potential role of these cells in T-cell development and immune function.

We thank Ronald Germain for discussions and critical review of the manuscript, Gina Stevenart for FACS analysis, Oberdan Leo and Jeffrey Bluestone for their gift of the 145-2C11 antibody, Lee Maloy for the TCR peptides and Brenda Marshall for excellent editorial assistance. We thank Richard Asofsky for his continued advice and support.

Note added in proof: A recent report by Budd *et al.* (*J. exp. Med.* 166, 577; 1987) describes a population of V β 8⁺ CD4⁺CD8⁺ thymocytes. The authors propose that these are immature T cells. In contrast, we conclude from the ontogeny studies and the phenotypic and functional analyses presented here that these cells are fully differentiated and represent a distinct T-cell lineage.

Received 22 May; accepted 2 July 1987.

- De la Hera, A., Toribio, M. L., Marquez, C. & Martinez-A., C. *Proc. natn. Acad. Sci. U.S.A.* **82**, 6268-6272 (1985).
- Bluestone, J. A., Pardoll, D. M., Sharrows, S. O. & Fowlkes, B. J. *Nature* **326**, 82-84 (1987).
- Lew, A. M. *et al. Science* **234**, 1401-1405 (1986).
- Pardoll, D. M. *et al. Nature* **326**, 79-81 (1987).
- Staerz, U. D., Rammensee, H.-G., Benedetto, J. D. & Bevan, M. J. *J. Immun.* **134**, 3994-4000 (1985).
- Fowlkes, B. J., Edison, L., Mathieson, B. J. & Chused, T. M. *J. exp. Med.* **162**, 802-822 (1985).
- Raulet, D. H., Garman, R. D., Saito, H. & Tonegawa, S. *Nature* **314**, 103-107 (1985).
- Snodgrass, H. R., Dembic, Z., Steinmetz, M. & von Boehmer, H. *Nature* **315**, 232-233 (1985).
- Samelson, L. E. *et al. Nature* **315**, 765-768 (1985).
- Leo, O., Foo, M., Sachs, D. H., Samelson, L. E. & Bluestone, J. A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 1374-1378 (1987).
- Haskins, K., Kubo, R., White, J., Pigeon, M., Kappler, J. & Marrack, P. *J. exp. Med.* **157**, 1149-1169 (1983).
- Roehm, N. *et al. Cell* **38**, 577-584 (1984).
- Crispe, L. N., Bevan, M. J. & Staerz, V. D. *Nature* **317**, 627-629 (1985).
- Kappler, J. W., Roehm, N. & Marrack, P. *Cell* **49**, 273-280 (1987).
- Behlke, M. A. *et al. J. exp. Med.* **165**, 257-262 (1987).
- Gunter, K. C., Malek, T. R. & Shevach, E. M. *J. exp. Med.* **159**, 716-730 (1984).
- Trowbridge, I. S., Lesley, J., Schulte, R., Hyman, R. & Trotter, J. *Immunogenetics* **15**, 399-412 (1982).
- Lesley, J. & Trowbridge, I. S. *Immunogenetics* **15**, 313-320 (1982).
- Dumont, F. J., Habberset, R. C., Coker, L. Z., Nichols, E. A. & Treffinger, J. A. *Cell Immunol.* **96**, 327-337 (1985).
- Takei, F., Secher, D. S., Milstein, C. & Springer, T. *Immunology* **42**, 371-378 (1981).
- Dumont, F. J., Coker, L. Z., Habberset, R. C. & Treffinger, J. A. *J. Immun.* **134**, 2357-2365 (1985).
- Malek, T. R., Robb, R. J. & Shevach, E. *Proc. natn. Acad. Sci. U.S.A.* **80**, 5694-5698 (1983).
- Morse, H. C. III, Davidson, W. F., Yetter, R. A. & Coffman, R. L. *Cell Immun.* **70**, 311-320 (1982).
- Murphy, E. D. in *Immunologic Defects in Laboratory Animals*, Vol. 2 (eds Gershwin, M. E. & Merchant, B.) 143-173 (Plenum, New York, 1981).
- Davidson, W. F., Dumont, F. J., Bedigian, H. G., Fowlkes, B. J. & Morse, H. C. III. *J. Immun.* **136**, 4075-4084 (1986).
- Steinberg, A. D., Roths, J. B., Murphy, E. D., Steinberg, R. T. & Ravache, E. S. *J. Immun.* **125**, 871-873 (1980).
- Ceredig, R., Lowenthal, J. W., Nabholz, M. & MacDonald, H. R. *Nature* **314**, 98-100 (1985).
- Hämmerling, G. J., Hämmerling, U. & Flaherty, L. *J. exp. Med.* **150**, 108-116 (1979).
- Ledbetter, J. A. & Herzenberg, L. A. *Immun. Rev.* **47**, 63-90 (1979).
- Manohar, V., Brown, E., Leiserson, W. M. & Chused, T. M. *J. Immun.* **129**, 532-538 (1982).
- Kruisbeek, A. M. *et al. J. exp. Med.* **161**, 1029-1047 (1985).

Functionally distinct subsites on a class II major histocompatibility complex molecule

Franca Ronchese, Ronald H. Schwartz & Ronald N. Germain

Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Mature T lymphocytes are activated by recognition of the combination of foreign protein antigen and membrane products of the major histocompatibility complex (MHC)¹. Studies of peptide antigen binding to detergent-solubilized class II MHC molecules (Ia) have established that peptide-Ia interaction occurs in the absence of the T-cell receptor and varies according to allele-specific features of Ia molecules²⁻⁴. The residues of immunogenic peptides thus contribute to two largely independent functions — the control of association with Ia molecules and the determination of the specificity of T-cell receptor binding^{5,6}. Two analogous and potentially independent functional sites have been postulated for Ia molecules⁷ — a region that controls binding to peptides and a region that interacts with T-cell receptors. Here we present evidence from functional analysis of recombinant class II molecules that these two postulated functional regions of Ia molecules do exist and can be independently manipulated, consistent with our recent demonstration of the segmental nature of Ia molecule structure-function relationships⁸.

To explore class II molecules for potentially independent functional regions, we attempted to map the residue(s) responsible for allele-specific variation in presentation of pigeon cytochrome *c* (PCC). B10.A (E β^k E α) mice are responders to PCC, whereas B10.A(5R)(E β^b E α) mice are nonresponders⁹. Both respond to cytochrome *c* from the tobacco hornworm moth (MCC)⁹. In both cases, the response is specific for a carboxy-terminal fragment of the cytochrome *c* molecule minimally encompassing residues 95-104¹⁰. It has also been shown that 90% of B10.A anti-PCC clones recognize both PCC and MCC

Table 1 Pattern of T-cell responses to transfectants expressing wild-type and recombinant E β genes

T cell	Antigen specificity	E β 1 origin			
		bb	bk	kb	kk
		Response to L cell transfectants			
Herpes I	Glycoprotein D of HSV, 1-23	+	+	+	+
Herpes II	Glycoprotein D of HSV, 1-23	+	+	+	+
G31	Staph. nuclease, 81-100	+	+	+	+
8I	λ repressor, 12-26	+	+	+	+
2B4	{ moth (86-90; 94-103)	+	+	+	+
AE7 cytochrome c		{ pigeon (81-104)	-	-	+
F1.A2					

+, Response of at least 50% of the response seen with parental E β E α -expressing L cells, based on a full antigen dose-response curve. -, Response undetectable even at high antigen concentrations. The indicated T-cell clones and hybridomas were cultivated in the presence of 5×10^4 L-cell transfectants and various concentrations of relevant antigen for 24 h in flat-bottom microtitre wells¹⁷. The response was determined in some cases by measuring either the ³H-TdR incorporation in an overnight pulse after culture for 48 h or by determining IL-2/IL-4 production after culture for 24 h as detailed in the legend to Fig. 2. The herpes simplex virus (HSV) glycoprotein D-specific hybridomas were obtained through the courtesy of Dr E. Heber-Katz, the nuclease specific clone G31 from Drs A. Finnegan and R. Hodes, and the λ repressor-specific hybridoma 8I from Drs J.-G. Guillet and M. Geffer.

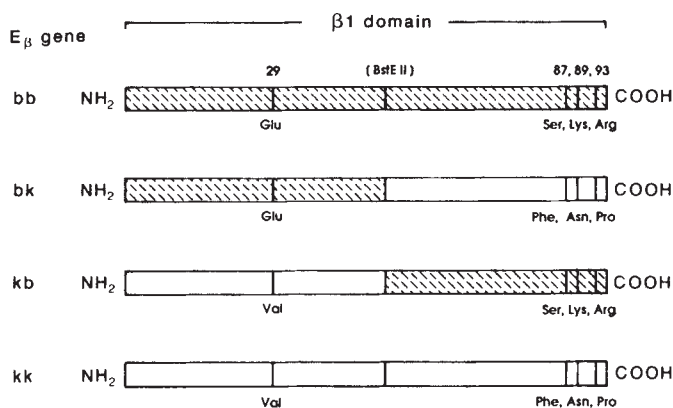


Fig. 1 Schematic diagram of the $\beta 1$ exons from wild type $E\beta^b$ and $E\beta^k$, and the recombinant $\beta 1$ exons of $E\beta^{bk}$ and $E\beta^{kb}$. The location of the *BstEII* restriction site with respect to polymorphic amino-acid residues in the encoded protein is shown.

Methods. An expressible genomic clone of $E\beta^b$ was constructed from the LS3/15 cosmid of Widera and Flavell¹⁶ by subcloning into a derivative of the plasmid vector pcEXV-1¹⁹ containing the bacterial XGPT gene under control of the Moloney long-terminal repeat (F.R. and R.N.G., unpublished data). A 5.8-kb *KpnI*/*EcoRI* fragment, comprising the $\beta 2$, TM and IC exons of $E\beta^b$ was blunt-ended and subcloned into the blunt-ended *EcoRI* site of pcEXV-gpt. The complete form of the gene was restored in a subsequent step by subcloning an 8.4-kb *Bam*HI fragment containing the L and $\beta 1$ exons into *Bam*HI digested DNA of the plasmid derived in the first subcloning step. A clone with the appropriate orientation of the two gene segments was identified by extensive restriction mapping. The final construct contains only two *EcoRI* restriction sites, 5' and 3' to the $\beta 1$ exon: these sites were used to exchange the $\beta 1$ exon of $E\beta^b$ for either the equivalent region of $E\beta^k$ (derived from cosmid H-2^k 7.1 of Steinmetz *et al.*²⁰) or for the $\beta 1$ regions from the *in vitro* recombinants $E\beta^{bk}$ or $E\beta^{kb}$. These recombinant $\beta 1$ exons were constructed by subcloning the *EcoRI* fragments containing the $\beta 1$ exons of $E\beta^b$ or $E\beta^k$ into pcEXV-1¹⁹, digesting the resulting plasmids with *BstEII* and *HindIII*, and exchanging the corresponding fragments from the two parental genes. In each final gene reconstruction, the leader exon, containing information for amino acids 1 to 6 in the mature protein, is derived from $E\beta^b$, which is identical to that of $E\beta^k$ (P. P. Jones, personal communication). These plasmids were individually transfected together with an expressible cDNA clone of $E\alpha$ (pcE α 7)¹⁹ into the DAP-3 subclone of murine L-cell fibroblasts and lines containing the transfected DNA selected using mycophenolic acid¹⁷. Stable clones expressing comparable levels of surface Ia for use in antigen-presentation assays were isolated by preparative cell sorting¹⁷.

with $E\beta^k E\alpha$, but only MCC with $E\beta^b E\alpha$ ¹¹⁻¹³. To explain these observations, it was hypothesized that the moth and pigeon cytochromes interacted in a distinct manner with the two Ia molecules, a process dependent on both the structure of the cytochrome *c* peptide and allelic variation in the class II molecules. Because T cells from both B10.A and B10.A(5R) mice could recognize MCC presented by accessory cells from either strain¹⁴, it also seemed that the two Ia molecules did not differ in their ability to fulfill the 'MHC-restriction' function for these particular T-cell receptors.

The $E\beta E\alpha$ molecules from the B10.A and B10.A(5R) strains are very similar: the $E\alpha$ chains are identical and the two $E\beta$ chains differ only in four amino-acid residues^{15,16}, position 29 in the second hypervariable region (HV2) and positions 87, 89 and 93 in the fourth hypervariable region (Fig. 1). To determine which of these residues was (were) critical to the failure of $E\beta^b E\alpha$ to present PCC, and to evaluate whether the ability to present PCC could be varied independently of the general MHC restriction properties of the Ia molecule, hybrid $E\beta$ molecules were generated in which position 29 from $E\beta^b$ was associated with residues 87, 89 and 93 from $E\beta^k$ and vice versa. The

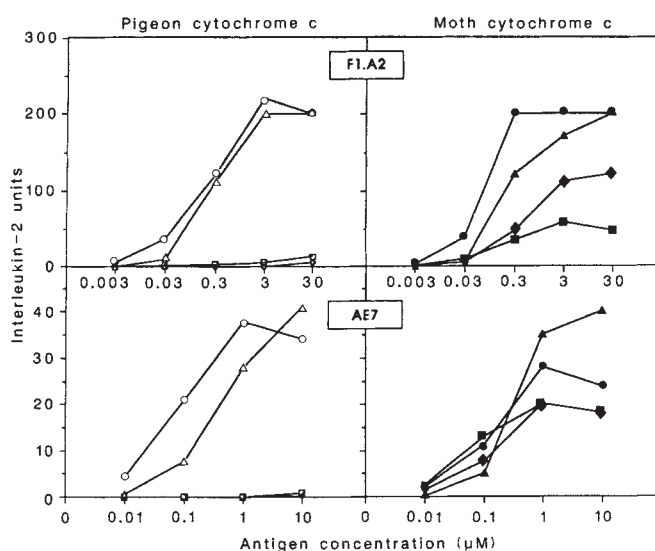


Fig. 2 Antigen dose-response curves for the cytochrome *c*-specific T-cell clones A.E7 and F1.A2; data are expressed as IL-2 units in response to various concentrations of cytochrome *c*. ($\diamond$, $\blacklozenge$) $E\beta^b E\alpha^-$; ($\circ$, $\bullet$) $E\beta^k E\alpha^-$; ($\square$, $\blacksquare$) $E\beta^b E\alpha^+$; ($\triangle$, $\blacktriangle$) $E\beta^k E\alpha^+$ -expressing L cell transfectants. Closed symbols, moth cytochrome *c* (86-90, 94-103); open symbols, pigeon cytochrome *c* (81-104). **Methods.** Cytochrome *c*-specific clones were maintained *in vitro* by cycles of stimulation with antigen and irradiated spleen cells, followed by a period of rest in the absence of antigen¹². Cells collected at the end of the resting phase were treated with a cocktail of I-A-specific and I-E-specific monoclonal antibodies and complement, and purified from dead cells by centrifugation over a Ficoll gradient. T cells (2×10^4) were cultured, in the presence of 5×10^4 mitomycin C-treated L-cell transfectants expressing $E\beta E\alpha$ molecules containing wild-type or recombinant $E\beta$ chains and various concentrations of soluble antigen, for 24 h in 96-well microtitre trays in 200 μ l final volume of RPMI-1640 supplemented with 10% fetal calf serum, glutamine, non-essential amino acids, 2-mercaptoethanol and gentamicin. The 81-104 antigenic fragment of pigeon cytochrome *c* was obtained by CNBr cleavage of the native protein and purified by gel-filtration chromatography²¹. The analogue of moth cytochrome *c* (moth residues 86-90; 94-103), kindly prepared by Dr Lee Maloy, was synthesized by the Merrifield solid-phase method and purified by HPLC²². After 18-24 h of incubation, T-cell culture supernatants were collected and their IL-2 content determined. CTL-L cells (5×10^3) were incubated for 24 h with six twofold dilutions of culture supernatant, pulsed for 6 h with 1 μ Ci of 3 H-TdR and harvested on glass fibre filters with a PHD automated cell harvester. IL-2 units were determined by interpolation and represent the supernatant dilution giving 50% of maximal proliferation compared to a reference IL-2 preparation.

wild-type or recombinant $E\beta$ genes were individually transfected together with $E\alpha$ into L cells, and lines suitable for antigen presentation assays were isolated.

As shown in Fig. 2, L cells expressing parental (wild-type) $E\beta E\alpha$ molecules stimulate T cells with the same pattern observed using splenic antigen-presenting cells. When L cells expressing Ia molecules involving recombinant $E\beta$ chains were analysed, it was found that the transfectants expressing molecules containing an $E\beta$ chain in which only residue 29 (Val) was from the *k* haplotype presented the PCC antigen to T cells in the same manner as wild-type $E\beta^k E\alpha$ transfectants, and, reciprocally, transfectants in which residue 29 in the expressed $E\beta$ chain was derived from the *b* haplotype (Glu) failed to present PCC. This was also seen using parental $E\beta^b E\alpha$ transfectants. Therefore, allelic variation at a single amino-acid residue determined the ability of an Ia molecule to present a specific antigen in a manner consistent with the *Ir* gene phenotype of the strain of origin.

To investigate the possible independence of this allele-specific

antigen-presentation property of the $E\beta E\alpha$ molecule from its general capacity to participate in MHC-restricted antigen recognition, additional T cells specific for various antigens in the context of $E\beta^k E\alpha$ were analysed. Surprisingly, despite their specificity for three antigens unrelated to cytochrome *c*, all of these T cells displayed the same MHC-restriction degeneracy seen with the MCC-specific T cells, that is, they all responded to antigen in the context of $E\beta^b E\alpha$, and also were stimulated by the Ia molecules involving the recombinant $E\beta$ chains (Table 1). Therefore, the structural difference between $E\beta^b$ and $E\beta^k$ responsible for the failure to present one specific antigen, PCC, does not affect the overall ability of the $E\beta E\alpha$ molecule to participate in antigen-specific, MHC-restricted T-cell stimulation. A simple explanation for this observation is that there are two functionally distinct sites on the Ia molecule, one controlling interaction with antigen and one for binding to that portion of the T-cell receptor involved in allele-related Ia recognition; only the former site appears to differ between $E\beta^b$ and $E\beta^k$.

These results stand in marked contrast to those obtained analysing the role of residues 67, 70 and 71 in the β_1 domain of $A\beta^b$ that constitute the *bm12* mutation. This work has shown that exchange of any one of the *b* residues for its *bm12* counterpart is sufficient to prevent the responses of most T cells¹⁷. It appears therefore that this site affects the recognition properties of the great majority of T cells, independent of their antigen specificity, as would be expected for a site that directly interacts with the T-cell receptor. Analysis of the sequences of allelic $E\beta$ chains reveals that the comparable region around the third hypervariable region (residues 60–75) is completely conserved between the $E\beta^b$ and $E\beta^k$ molecules, but differs from other alleles (such as $E\beta^s$, ref. 18) that can present PCC and MCC to B10.A-derived T cells in a much smaller number of cases^{12,13}. This suggests that a critical region involved in formation of the MHC-restriction site is the third HV region of the β chain, and that the contribution of this region to T-cell stimulation can be functionally independent of the effects of changes in HV2 that affect antigen presentation.

These distinct roles of different segments of the $E\beta$ chain are likely to depend primarily on distinct spatial locations of the hypervariable regions in the mature folded Ia molecule. Analysis of the $A\beta A\alpha$ molecule⁸ has shown that the amino-terminal half of the $A\beta$ first domain (encompassing HV1 and HV2) is mainly involved in determining $\alpha\beta$ chain interactions and I-A conformation, whereas the carboxy-terminal half (encompassing HV3 and HV4) is the site primarily recognized by monoclonal antibodies. This suggests that the region of the β chain containing HV3 is clearly exposed on the surface where it is available for direct interaction with antibodies and T-cell receptors. This exposed position could account for the importance of this region of allelic variation in the process of MHC-restricted antigen recognition by most T cells.

The precise physical basis for the effect of alterations in residue 29 of the $E\beta$ chain on presentation of PCC has not been directly determined by our experiments. Buus and colleagues⁴ have recently reported a 10–15-fold difference in the binding affinity of PCC peptide to $E\beta^k E\alpha$ as compared to $E\beta^b E\alpha$. Given the complete correlation of our functional studies with these binding data, it appears reasonable to assume that the Val-Glu interchange at position 29 controls the efficiency of PCC binding to the $E\beta^b E\alpha$ and $E\beta^k E\alpha$ molecules, and hence the availability of antigen-Ia ligand for stimulation via the T-cell receptor. If this is confirmed by direct testing, it would demonstrate the ability of a single amino acid to control immune response potential by changing the extent to which antigen associates with class II Ia molecules, without significantly affecting the general function of that Ia molecule as an MHC-restriction element.

We wish to thank Drs A. Finnegan, M. Geftter, J. G. Guillet, E. Heber-Katz and R. Hodes for providing T-cell clones and hybridomas; Drs M. Steinmetz and G. Widera for providing

cosmid clones; Dr Jim Miller for providing the $E\alpha$ clone; Dr B. Fox for critically reading the manuscript; and Ms Shirley Starnes for editorial assistance.

Received 21 May; accepted 30 July 1987.

1. Schwartz, R. H. *A. Rev. Immun.* **3**, 237–261 (1985).
2. Babbitt, B. P., Allen, P. M., Matsueda, G., Haber, E. & Unanue, E. R. *Nature* **317**, 359–361 (1985).
3. Buus, S., Sette, A., Colon, S. M., Jenis, D. M. & Grey, H. M. *Cell* **47**, 1071–1077 (1986).
4. Buus, S., Sette, A., Colon, S. M., Miles, C. & Grey, H. M. *Science* **235**, 1353–1358 (1987).
5. Hansburg, D., Fairwell, T., Schwartz, R. H. & Appella, E. *J. Immun.* **131**, 319–324 (1983).
6. Allen, P. M. *et al. Nature* **327**, 713–715 (1987).
7. Heber-Katz, E., Hansburg, D. & Schwartz, R. H. *J. molec. Cell. Immun.* **1**, 3–14 (1983).
8. Braunstein, N. S. & Germain, R. N. *Proc. natn. Acad. Sci. U.S.A.* **84**, 2921–2925 (1987).
9. Solinger, A. M., Ultee, M. E., Margoliash, E. & Schwartz, R. H. *J. exp. Med.* **150**, 830–848 (1979).
10. Carbone, F. R., Fox, B. S., Schwartz, R. H. & Paterson, Y. *J. Immun.* **138**, 1838–1844 (1987).
11. Heber-Katz, E. *et al. J. exp. Med.* **155**, 1086–1099 (1982).
12. Hedrick, S. M. *et al. Cell* **30**, 141–152 (1982).
13. Matis, L. A. *et al. J. Immun.* **130**, 1527–1535 (1983).
14. Ashwell, J. D. & Schwartz, R. H. *Nature* **320**, 176–179 (1986).
15. Mengle-Gaw, L. & McDevitt, H. O. *Proc. natn. Acad. Sci. U.S.A.* **80**, 7621–7625 (1983).
16. Widera, G. & Flavell, R. A. *EMBO J.* **3**, 1221–1225 (1984).
17. Ronchese, F., Brown, M. A. & Germain, R. N. *J. Immun.* **139**, 629–638 (1987).
18. Mengle-Gaw, L. & McDevitt, H. O. *Proc. natn. Acad. Sci. U.S.A.* **82**, 2910–2914 (1985).
19. Miller, J. & Germain, R. N. *J. exp. Med.* **164**, 1478–1489 (1986).
20. Steinmetz, M. *et al. EMBO J.* **3**, 2995–3003 (1984).
21. Corradin, G. P. & Harbury, H. A. *Biochim. biophys. Acta* **221**, 489–496 (1970).
22. Suzuki, G. & Schwartz, R. H. *J. Immun.* **136**, 230–239 (1986).

Predictable acquisition of a new MHC recognition specificity following expression of a transfected T-cell receptor β -chain gene

Takashi Saito & Ronald N. Germain

Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Activation of mature T lymphocytes requires specific corecognition of antigen together with membrane-associated glycoprotein products of the major histocompatibility complex (MHC)¹. This dual specificity is determined by a single receptor structure consisting of a clone-specific $\alpha\beta$ heterodimer^{2–8}. Because both the α and β subunits possess unique combining-site-containing V regions⁹, it remains an open issue as to what contribution each of the two chains of the receptor makes to the antigen versus MHC recognition specificities of the complete dimer present on any given T cell or in the T-cell pool as a whole. In the present work, we have used DNA-mediated gene transfer to express a new α or β chain in a recipient murine T-cell hybridoma possessing a related antigen but distinct MHC specificity compared to the receptor-gene donor. Our results demonstrate that a β -gene transfected hybridoma expresses new receptors with a predictable hybrid specificity, establishing that the β chain has the predominant role in MHC molecule recognition in this model.

A minimal requirement for successful stimulation of a T cell is the formation of an appropriate number of antigenic peptide-Ia molecule complexes. Previous studies using somatic T cell hybrids^{10,11} to investigate the molecular basis for antigen versus MHC molecule recognition did not explicitly take into account the recently described allele-specific interactions of immunogenic peptides and Ia molecules^{12–15}. Thus, the failure of certain antigen-Ia pairs to form may have prevented detection of new combinations of receptor specificities in such experiments. To ensure that after gene transfer we would not fail to detect new receptor specificities due to a lack of such ligand complexes, we have used a model system in which the ability of a particular antigen (moth cytochrome *c*, MCC) to form immunogenic complexes with either of two distinct Ia molecules ($E\beta^k E\alpha$ and $E\beta^b E\alpha$) has already been established¹⁵. Two well-

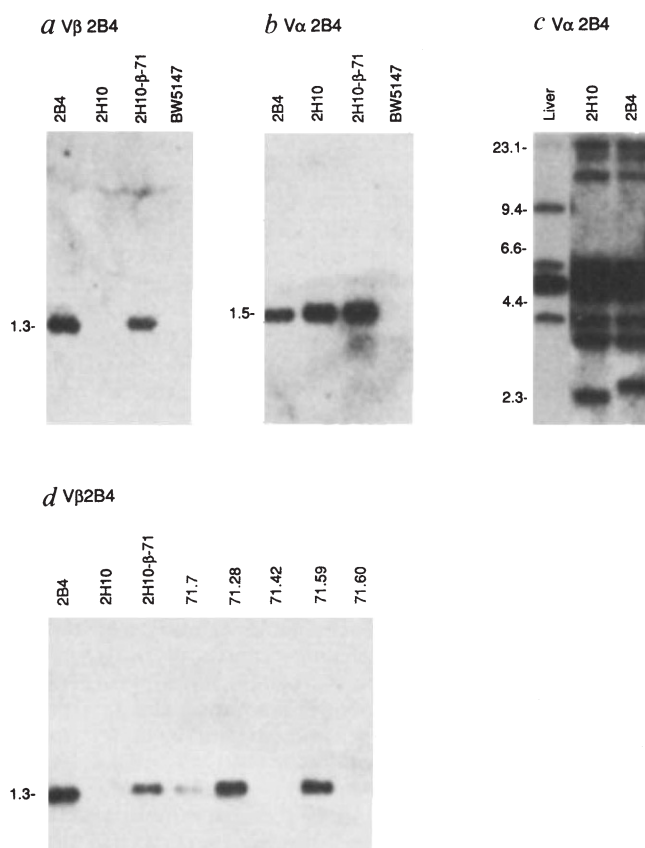


Fig. 1 Northern (*a*, *b* and *d*) and Southern blot analysis (*c*) of 2B4, 2H10 and 2B4 β -transfected 2H10 clones.

Methods. 2B4 β -expressing 2H10 clones represented by 2H10- β -71 were obtained by transfection of 2H10 with p2B4 β SV2gpt by protoplast fusion as previously described⁸. The clones were selected in the presence of 0.3 $\mu\text{g ml}^{-1}$ mycophenolic acid, 250 $\mu\text{g ml}^{-1}$ xanthine and 15 $\mu\text{g ml}^{-1}$ hypoxanthine. Various 2B4 β gene expression variants were derived from 2H10- β -71 by culture in the absence of mycophenolic acid, followed by cloning by limiting dilution. Screening was carried out by RNA dot blot using a 2B4 V β probe. For Northern blot analysis, total cytoplasmic RNA (10 μg) isolated from 2B4, 2H10 or 2B4 β -transfected 2H10 clones was electrophoresed in a 1% agarose gel containing 0.22 M formaldehyde, and transferred to nitrocellulose²¹. For Southern blot analysis, DNA (15 μg) from B10.A liver, 2B4 or 2H10 was digested with *Hind*III, electrophoresed on a 0.8% agarose gel, and transferred to nitrocellulose²². Blots were hybridized with random hexamer-primed, ³²P-labelled probes²³, washed and autoradiographed. The probes used were: *a* and *d*, a 310-base-pair (bp) *Eco*RI-*Sca*I fragment corresponding to 2B4 V β (V β 3, ref. 24) prepared from p2B4 β SV2gpt (ref. 8), *b* and *c*, a 340-bp *Pst*I-*Pvu*II fragment corresponding to 2B4 V α (V α 11.2, ref. 25) from a pGEM-2B4 α ⁸. The relative sizes (in kb) of the bands are indicated in the margins.

characterized T-cell hybridomas¹⁶ specific for MCC, but differing in their Ia recognition specificities (2B4 and 2H10) were chosen for these experiments. These T cells respond to both pigeon cytochrome *c* (PCC) and MCC when presented by E β^k E α -bearing antigen-presenting cells (APC). 2B4 responds to MCC but not PCC on E β^k E α -bearing APC, whereas 2H10 does not respond to either antigen presented by E β^k E α -bearing APC¹⁶. As shown in Fig. 1*a*, the receptors of 2B4 and 2H10 use completely distinct V β segments, as a 2B4 V β probe does not hybridize to any messenger RNA in 2H10. A 2B4 V α probe does detect normal sized α -chain transcripts in 2H10 (Fig. 1*b*), consistent with the previous reports of a very high frequency of V α 11 gene family use in cytochrome *c* specific T cells^{17,18}. But Southern blot analysis shows two distinct bands of 2.4 kilobases

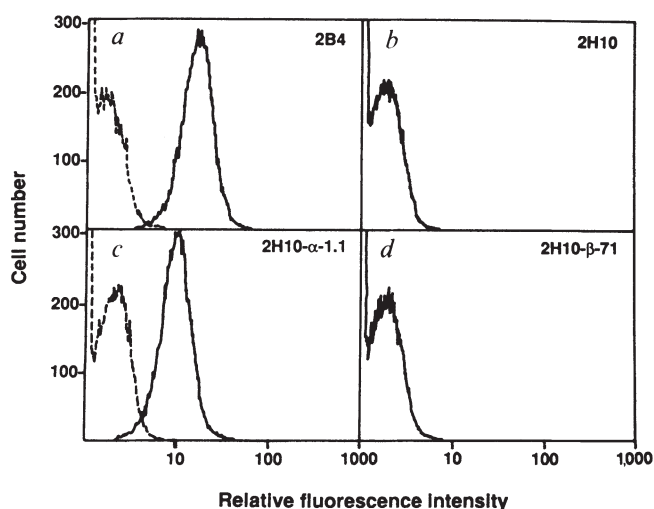


Fig. 2 Cell surface staining profiles of *a*, 2B4, *b*, 2H10, *c*, 2B4 α -chain and *d*, 2B4 β -chain gene-transfected 2H10 clones. The 2B4 α -chain-expressing 2H10 clone 2H10- α -1.1 was derived by transfection of 2H10 with p2B4 α Fneo⁸, followed by screening for surface staining with A2B4.2 antibody. Aliquots of 5×10^5 2B4, 2H10 2H10- α -1.1 or 2H10- β -71 cells were stained with A2B4.2 culture supernatant (monoclonal anti-2B4 α antibody) (—) or the mouse myeloma protein MOPC 104E (---) as a negative control. After incubation at 4 °C for 30 min, the cells were washed and incubated for an additional 30 min with fluorescein-conjugated goat anti-mouse immunoglobulin, then washed again. The stained cells were analysed on a fluorescence-activated cell sorter. The results shown are an analysis of 10^4 cells.

(kb) and 2.6 kb, respectively, using the V α 2B4 probe on *Hind*-III-digested DNA from these cells (Fig. 1*c*). The data of Sorger *et al.*¹⁹ indicate that these different rearrangements probably reflect the use by these two hybridomas of distinct J segments and/or V α 11 family members. This is also consistent with the failure of 2H10 to react with the A2B4 anti-clonotypic monoclonal antibody⁵, which we have previously shown to be specific for the 2B4 α chain⁸. Thus, these two T cells use distinct, though related, $\alpha\beta$ receptors to mediate their antigen and MHC molecule specificities.

Recombinant plasmid constructs containing expressible forms of the 2B4 α -chain or β -chain receptor genes⁸ were transfected into 2H10 using protoplast fusion. The transfectants (for example, 2H10- α -1.1) containing the 2B4 α -chain gene were analysed by flow microfluorimetry and shown to have significant cell surface expression of new receptors containing the 2B4 α chain (Fig. 2*c*). Because no antibody was available for the 2B4 β chain, 2B4 β -gene transfectants were analysed for 2B4-specific mRNA using a V β probe. 1.3-kb mRNA hybridizing to this probe was detected in most of the β -gene transfectants, as shown for a representative clone (2H10- β -71) in Fig. 1*a*, lane 3. A series of 2B4 β -chain expression variants was then derived from 2H10- β -71 and the transcription of the transfected gene was analysed in these subclones by Northern blotting (Fig. 1*d*). Whereas 2H10- β -71.28 (71.28) and 71.59 have high levels of 2B4 β -chain mRNA, 71.7 has a low level, and 71.42 and 71.60 no detectable amount of this transcript. These clones could thus be used to correlate changes in antigen or MHC molecule specificity with the presence or absence of 2B4 β -chain mRNA, and hence, with the presumptive presence of new 'hybrid' T-cell receptors involving the transfected β -chain.

Each β -gene transfected clone was stimulated with the MCC peptide analogue, DASP, in the presence of either B10.A (5R) (E β^k E α) or B10.A (E β^b E α) antigen-presenting cells and lymphokine production measured (Fig. 3*a, b*). The shift in the response profiles of the 2B4 β -expressing 2H10 transfectant subclones towards the 2B4 pattern is consistent with the presence

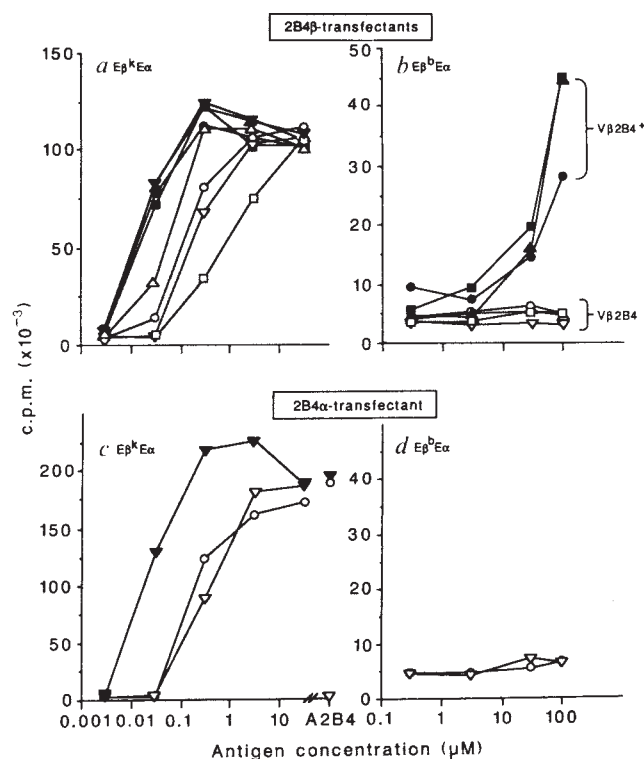


Fig. 3 Lymphokine responses of 2B4 β (a and b) or α (c and d) gene-transfected 2H10 clones to moth cytochrome *c* peptide analogue DASP¹ in association with B10.A (a, c) or B10.A(5R) (b, d) antigen-presenting cells. T cells (5×10^4) were cultured in the presence of 5×10^5 irradiated B10.A (E β^b E α) or B10.A(5R) (E β^b E α) spleen cells and the indicated dose of DASP. After cultivation for 24 h, 1:4 dilutions of culture supernatants were tested for lymphokine (IL-2) using the IL-2-dependent cell line CTLL as described previously⁸. Each point represents the mean of triplicate cultures. Symbols used are: a and b, 2H10- β -71 (■), 71.28 (●), 71.59 (▲), 71.7 (△), 71.42 (□), 71.60 (○), 2B4 (▼) and 2H10 (▽); c and d, 2H10- α -1.1 (○), 2B4 (▼) and 2H10 (▽). On the right side of panel c, points indicate the responses of 2B4 (▼), 2H10 (▽), and 2H10- α -1.1 (○) to immobilized A2B4.2 antibody.

of new 2B4 β -containing heterodimers on these cells. More importantly, the three 2H10 transfectants expressing high levels of 2B4 β -chain mRNA, and, presumably, significant amounts of a 'hybrid' T cell receptor (71, 71.28, and 71.59), responded to DASP presented by B10.A(5R) spleen cells, showing an MHC recognition pattern absent in the parent 2H10 hybridoma, but present in the gene donor. The clones with little or no 2B4 β -chain expression failed to show this gain in specificity, consistent with the change in phenotype in the E β^b E α -responsive transfectants being associated with the expression of receptors containing the transfected gene product.

To determine whether the predictable change in MHC molecule specificity seen in recipients of the 2B4 β chain gene was associated only with this chain, the function of 2B4 α -transfectants was analysed. As shown in Fig. 3c and d, despite the expression of high levels of 2B4 α -containing receptors on the cell surface, these transfected T-cell hybridomas did not gain the capacity to respond to DASP and E β^b E α . The new receptors were functional, as the cells bearing them could be stimulated to produce lymphokine by immobilized anti-clonotypic antibody (Fig. 3c). These results establish the special role of the 2B4 β -chain in altering the MHC molecule specificity of 2H10.

The gene transfer experiments presented here were designed to explore the functional contribution of each clonotypic receptor chain to T-cell specificity and to look for any possible

independence of the two variable regions with respect to antigen and MHC molecule recognition. Our results provide two new pieces of information concerning the molecular basis of T-cell dual recognition. First, the acquisition of reactivity to DASP plus E β^b E α following transfer and expression of the 2B4 β -chain but not the 2B4 α -gene in the 2H10 hybridoma provides direct evidence for a major role of the receptor β -chain in control of MHC molecule specificity in the cytochrome *c* model system, in agreement with sequence analysis experiments¹⁷⁻¹⁹. Second, by using α -chain and β -chain genes from two T cells that were independently selected by antigen and MHC molecules, it has been possible to construct a hybrid receptor whose functional specificity is that which is expected based on the known reactivities of the parent T lymphocytes. This creation of a functional 'mixed-specificity' receptor raises the possibility that the two variable segments of the $\alpha\beta$ heterodimer make relatively independent contributions to recognition of MHC molecules and physically associated antigen.

Clearly, no single example of this type, even if it were to reflect a strict division between the α - and β -chains of binding sites for the specific antigen and MHC molecule in question, could be generalized to all T-cell receptors or antigen-MHC molecule combinations. It also remains possible that the β -chain only sterically limits antigen-plus-Ia recognition mediated by the relatively conserved α -chain present in most cytochrome *c*-specific T cells. Nonetheless, we feel that our ability to generate a functional hybrid receptor under conditions taking the nature of antigen-MHC molecule interactions into account is highly relevant to the debate concerning the structural basis for T-cell dual recognition. This positive result indicates a need for caution in accepting negative data as strong evidence for a lack of any unique or independent role of the α chain or β chain in mediating the dual specificity of T cells.

In fact, our findings, together with the recent report of a striking association of a newly identified germline V β segment with receptor specificity for E β E α molecules²⁰, raise the question of whether a simple analogy between the T-cell receptor and immunoglobulin with respect to combining site construction is appropriate. T-cell receptor genes may have been selected during the course of evolution such that the two sets of variable regions have distinct roles in the dual recognition process that distinguished T lymphocytes from B lymphocytes. Resolution of this issue will clearly require not only additional functional experiments of the type reported here, but direct binding analysis using isolated receptor chains, and ultimately, crystallographic studies on the trimolecular complex of T-cell receptor, antigen, and MHC molecule.

We thank R. H. Schwartz and L. E. Samelson for providing 2B4 and 2H10 cells and A2B4 antibody, respectively, S. Starnes for preparation of the manuscript and J. A. Bluestone, R. H. Schwartz, E. M. Shevach and A. Singer for reviewing the manuscript.

Received 21 May; accepted 30 July 1987.

1. Schwartz, R. H. *A. Rev. Immun.* **3**, 237-261 (1985).
2. Allison, J. P., McIntyre, B. W. & Bloch, D. J. *Immun.* **129**, 2293-2300 (1982).
3. Meuer, S. C. *et al. J. exp. Med.* **157**, 705-719 (1983).
4. Haskins, K. *et al. J. exp. Med.* **157**, 1149-1169 (1983).
5. Samelson, L. E., Germain, R. N. & Schwartz, R. H. *Proc. natn. Acad. Sci. U.S.A.* **80**, 6972-6976 (1983).
6. Yague, J. *et al. Cell* **42**, 81-87 (1985).
7. Dembic, Z. *et al. Nature* **320**, 232-238 (1986).
8. Saito, T., Weiss, A., Miller, J., Norcross, M. A. & Germain, R. N. *Nature* **325**, 125-130 (1987).
9. Kronenberg, M., Siu, G., Hood, L. E. & Shastri, N. A. *Rev. Immun.* **4**, 529-591 (1986).
10. Kappler, J. W., Skidmore, B., White, J. & Marrack, P. *J. exp. Med.* **153**, 1198-1214 (1981).
11. Haas, W., Mathur-Rochat, J., Kisielow, P. & von Boehmer, H. *Eur. J. Immun.* **15**, 963-965 (1985).
12. Babbitt, D. P., Allen, P. M., Matsueda, G., Haber, E. & Unanue, E. R. *Nature* **317**, 359-361 (1985).
13. Buus, S., Alessandro, S., Colon, S. M., Jenis, D. M. & Grey, H. M. *Cell* **47**, 1071-1077 (1986).
14. Guillet, J. G. *et al. Science* **235**, 865-870 (1987).
15. Buus, S., Sette, A., Colon, S. M., Miles, C. & Grey, H. M. *Science* **235**, 1353-1358 (1987).
16. Hedrick, S. M. *et al. Cell* **30**, 141-152 (1982).
17. Fink, P., Matis, L. A., McElligott, D. L., Bookman, M. A. & Hedrick, S. M. *Nature* **321**, 219-226 (1986).
18. Winoto, A. *et al. Nature* **324**, 679-682 (1986).

19. Sorger, S. B., Hedrick, S. M., Fink, P. J., Bookman, M. A. & Matis, L. A. *J. exp. Med.* **165**, 279-301 (1987).
20. Kappler, J. W. *et al.* *Cell* **49**, 263-271 (1987).
21. Thomas, P. *Proc. natn. Acad. Sci. U.S.A.* **77**, 5201-5205 (1980).
22. Southern, E. M. *J. molec. Biol.* **98**, 503-517 (1975).
23. Feinberg, A. P. & Vogelstein, B. *Analyt. Biochem.* **132**, 6-13 (1983).
24. Barth, R. K. *et al.* *Nature* **316**, 517-523 (1985).
25. Arden, B., Klotz, J. L., Siu, G. & Hood, L. *Nature* **316**, 783-787 (1985).

Frequent *c-fms* activation by proviral insertion in mouse myeloblastic leukaemias

Sylvie Gisselbrecht*, Serge Fichelson*, Brigitte Sola*,
Didier Bordereaux*, Annie Hampe†, Catherine André†,
Francis Galibert† & Pierre Tambourin*

*INSERM U 152, CNRS UA 628, Hôpital Cochin,
27 rue du Faubourg Saint Jacques, 75014 Paris, France
†Laboratoire d'Hématologie Expérimentale, CNRS-LOI,
Hôpital Saint-Louis, 75010 Paris, France

Retroviruses lacking oncogenes can induce tumours in animals, and the tumour cells are frequently found to contain proviral DNA inserted next to a proto-oncogene, which is thus placed under the regulatory control of the retroviral long terminal repeat (LTR). This altered regulation leads to overexpression of the proto-oncogene, which presumably contributes to the growth properties of the tumour cells¹. *fim-2* has been described as a retroviral integration site frequently and specifically involved in murine myeloblastic leukaemias induced *in vivo* or *in vitro* by the replication-competent Friend murine leukaemia virus (F-MuLV)². Here we report that *fim-2* spans the 5'-end of the murine proto-oncogene *c-fms*, known to code for a transmembrane glycoprotein with tyrosine kinase activity probably identical to the receptor of the haemopoietic growth factor, monocyte-macrophage colony-stimulating factor (M-CSF or CSF-1)³. Proviral integration in the *fim-2* region results in a high expression of a normal sized *c-fms* messenger RNA. We also observe that some tumours have lost the *fim-2/c-fms* germ line allele. These results provide the first evidence for the presumed involvement of *c-fms* in myelomonocytic leukaemias.

The molecular cloning and the genetic map of two F-MuLV integration sites in murine myeloblastic leukaemia have been previously described². These two regions, *fim-1* and *fim-2*, were cloned from a genomic library of a myeloblastic cell line harbouring five integrated F-MuLV proviruses. These viral integration sites appeared to be specifically involved in murine myeloblastic transformation, as no DNA rearrangement of these sites was found in 20 lymphoid or erythroid leukaemias induced by the same virus². Other murine myeloblastic leukaemias have been examined and out of 68 myeloblastic cell line or tumour DNAs, 14 (20%) displayed proviral insertion in the *fim-2* region. Among these 14 myeloblastic leukaemias, seven were *in vivo*

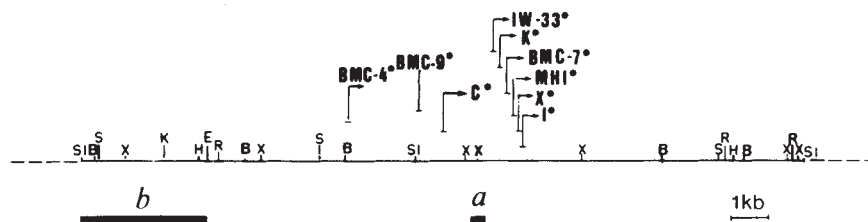
primary leukaemias, and seven cell lines derived from *in vitro* F-MuLV-infected long-term bone marrow cell cultures^{4,5}. Out of these 14 rearranged cell line or tumour DNAs, nine have been analysed in detail. Figure 1 shows that the proviruses integrated in the *fim-2* region all had the same orientation and fell in a cluster of 5 kilobases (kb). Provirus inserts in the *fim-2* locus were either full-length or deleted F-MuLV genomes, one of the inserts being only 0.9-kb long and corresponding to a single LTR structure. Interestingly, 3 out of 14 cell lines (C, BMC-9, K) lacked the *fim-2* germ line allele (ref. 2 and data not shown).

To examine whether this integration site corresponded to an expressed coding gene, we looked for RNA expression using different probes derived from overlapping phages. One of these probes (255-ES34), referred to as probe *b* (Fig. 1), was located 4 kb away from the viral integration cluster and detected a major 3.8-kb RNA species. This mRNA was found in all tested myeloblastic leukaemic cells exhibiting a *fim-2* rearranged allele (6/6) (see Fig. 2A, lanes *b*, *c*, *f*, *g*, *h*). In 11 other myeloblastic cell lines or tumours in which no *fim-2* rearrangement had so far been detected, three cell lines showed a high *fim-2* mRNA expression (see for example, Fig. 2A, lanes *e*, *i*). This expression could be due to unidentified F-MuLV integration in other regions of the *fim-2* gene further away from the major cluster of proviral integration, as found in other systems⁶⁻⁹. Two of the 11 cell lines did not express *fim-2* mRNA (Fig. 2A, lanes *d*, *j*) and six others, including the murine myelomonocytic WEHI-3B cell line, showed weak expression of the same sized mRNA (for example, lanes *a*, *j*, *k*, Fig. 2A). No expression was detected in five F-MuLV-induced erythroleukaemic cell lines, three plasmacytomas, four Abelson-induced pre-B cell lines and three thymomas, suggesting specific expression of *fim-2* in myeloblastic leukaemias.

The functional activation of cellular oncogenes by proviral insertion has been shown in several systems¹. We have studied in particular two cellular oncogenes encoding mRNA species similar in size to the *fim-2* mRNA, namely *c-myb*¹⁰ and *c-fms*¹¹. Viral insertion within the *c-myb* locus has already been described in some murine myeloblastic cell lines^{8,9}. We found that *c-myb* was uniformly expressed in myeloblastic cells and in erythroleukaemias in which no *fim-2* mRNA was detected (data not shown), whereas the pattern of *c-fms* expression was identical to that observed for *fim-2* (Fig. 2A and B).

To establish the possible relationship between *fim-2* and *c-fms*, we sequenced 270 base pairs (bp) of non-repetitive sequences derived from probe *b*. Considerable homology (81%) was found between this sequence and 5'-noncoding sequences of the human *c-fms* gene (Fig. 3b). We also cloned the human counterpart of the murine *fim-2* gene, and sequenced 900 bp derived from a 3.2-kb *EcoRI* fragment which hybridized with the mouse probe *b* and was free of repetitive elements (Fig. 3a). This sequence was found to be identical to a 5'-portion of human *c-fms* that included the second and third coding exons and an intervening intron (A. Hampe *et al.*, unpublished data). The human probe revealed exactly the same pattern of expression as the mouse probe *b* when used on mouse mRNA (not shown). From these

Fig. 1 Restriction map of the *fim-2* domain. Restriction endonuclease sites were determined from two overlapping phage clones obtained from a myeloblastic cell line genomic library². This map was confirmed by Southern blot analysis of mouse DNA. Symbols are as follows: E, *EcoRI*; R, *EcoRV*; H, *HindIII*; B, *BamHI*; S, *SacI*; X, *XbaI*; K, *KpnI*; Sl, *Sall* (the two extreme *Sall* sites are the phage cloning sites). Black boxes drawn under the map represent the probes used. Probe *a* is a 0.4-kb *PvuII-BstEII* fragment used to search for DNA rearrangements². Probe *b* is a 3.4-kb *Sall-EcoRI* fragment used in Northern blots. The location and orientation of F-MuLV proviruses in myeloblastic leukaemias (represented by *) are indicated by arrows. The orientation of the F-MuLV in the BMC-9 cell line was not determined because it contained only one LTR.



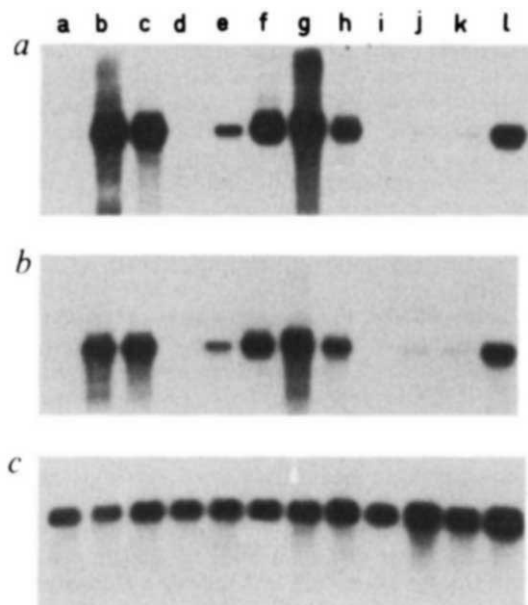


Fig. 2 Northern blot analysis of myeloblastic leukaemia poly(A)⁺ RNAs. RNAs were prepared using the hot phenol technique²⁴ and poly(A)⁺ RNAs selected on oligo(dT) columns. RNA (2 µg) was electrophoresed after glyoxal denaturation, blotted on to Gene Screen membrane, processed, hybridized, washed and dehybridized as recommended by the supplier (New England Nuclear). The same blot was analysed successively with different ³²P nick-translated probes (3 × 10⁸ c.p.m. per µg) under stringent conditions. *a*, Probe *b* of the *fin-2* region described in Fig. 1. *b*, A 1-kb *Pst*I–*Bgl*III fragment derived from *v-fms*²⁵. *c*, Probe RGAPDM or rat glyceraldehyde-3'-phosphodehydrogenase²⁶ used to ensure uniform loading. Lanes *b*, *c*, *f*, *g*, *h*: cell lines rearranged at the *fin-2* locus (*b*: BMC-9, *c*: I, *f*: BMC-4, *g*: C, *h*: X). Lanes *a*, *d*, *e*, *i*, *j*–*l*: cell lines not rearranged at the *fin-2* locus (*a*: WEHI-3B, *d*: BMC-61, *e*: MX1, *i*: NAD7, *j*: MA11, *k*: ADB22, *l*: A201-33). Most of these cell lines have been previously described^{2,4,5}.

results, it is concluded that *fin-2* spans the 5'-end of the mouse *c-fms*. Therefore only this region of the murine *fin-2/c-fms* gene, probably comparable in size to its human counterpart (35 kb), has been investigated using the *fin-2* probes. It is unlikely that a proviral integration at the 3'-end of this gene would be detected. Consequently, high mRNA expression in cell lines without detectable *fin-2* rearrangements could possibly be explained by viral integration either at the *c-fms* 3'-end or further upstream 5' to the *c-fms*.

F-MuLV integration of the *fin-2/c-fms* region was associated with a high production of *c-fms* mRNA, probably due to enhancer elements present in the proviral LTR. This hypothesis is supported by the following observations: first, the orientation of integrated proviruses at the *fin-2/c-fms* locus is head-to-head in relation to the 5'-region of the murine *c-fms*; second, this integration does not seem to affect the size of the *fin-2/c-fms* mRNA; third, the cell line with only one LTR structure produces large amounts of *c-fms* mRNA (Fig. 2A; lane *b*) and fourth, cell lines lacking the germ line allele also strongly expressed *c-fms* mRNA.

The *v-fms* oncogene present in the McDonough strain of the feline sarcoma virus (SM-FeSV)¹² encodes the complete ligand-binding domain of the CSF-1 receptor¹³, but *v-fms* is constitutively active as a kinase and displays no change in enzyme activity when it binds CSF-1¹⁴. SM-FeSV does not induce leukaemias but causes fibrosarcomas in cats^{12,15}. In humans, *c-fms* has been mapped to chromosome 5q33.3–34^{16,17} and is located near to the CSF-1 locus at 5q33.1¹⁸. No specific chromosomal rearrangement involving this gene has been associated with leukaemias, although the loss of either the entire chromo-

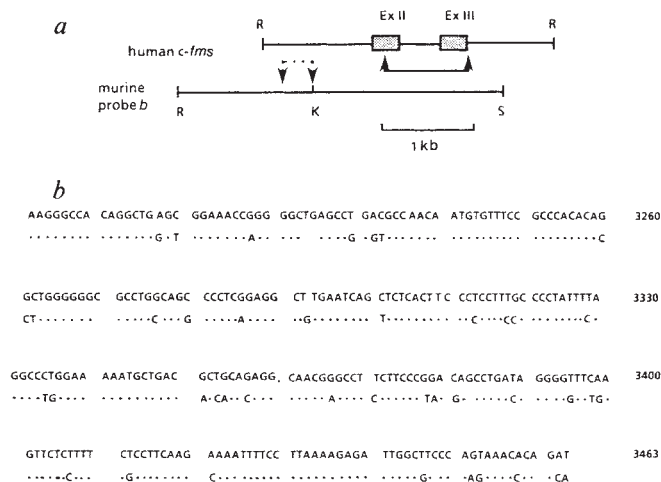


Fig. 3 *a*, Schematic representation of the human *fin-2/H6RI* fragment and the murine probe *b*. Dot blots of human *c-fms* genomic phage clones (kindly provided by C. Sherr²⁷) were probed using the human *fin-2/H6RI* fragment. Clone 11, corresponding to the 5' region of the human *c-fms* gene, strongly hybridized with the human *fin-2/H6RI* probe. 900 bp from the *fin-2/H6RI* fragment (solid line delimited by the arrows) were sequenced and found to be identical to a 5' portion of human *c-fms* that included half of the second coding exon (exII), an intervening sequence and the third coding exon (exIII). The 5'-coding exon (exon I) was identified previously¹⁴ and included sequences encoding a portion of the 5'-untranslated region of *c-fms* mRNA as well as the initiation codon and signal peptide of the *c-fms* gene product²⁸. *b*, Sequence comparison between the human *c-fms* gene and the mouse probe *b*. A 270-bp sequence from the mouse probe *b* (dotted line delimited by arrows) were found to have 81% homology with part of the human *c-fms* gene. These data allowed us to align the *fin-2* probe *b* with the corresponding human *c-fms* region. Position numbers for human *c-fms* were defined according to the first AUG initiation codon. Dots indicate nucleotides identical in mouse and human *c-fms*. Gaps have been introduced to allow optimal alignment. Symbols: R, *Eco*RI; K, *Kpn*I; S, *Sal*I.

Methods. A genomic library was obtained from normal human lymphocytes as previously described²⁹. High molecular weight DNA was partially digested with *Sau*III restriction enzyme and ligated to EMBL4 phage arms. 2.5 × 10⁶ independent recombinant phages were obtained and 10⁶ were screened with the mouse probe *b* described in Fig. 1. Five positive clones were isolated and purified by 3 successive rounds of cloning. One of them, named *fin-2/H6*, was mapped using 8 restriction enzymes. The *fin-2/H6* restriction fragments which hybridized to the mouse probe *b* had the size predicted from Southern blot analysis of human genomic DNA (not shown). A 3.2-kb *Eco*RI fragment (*fin-2/H6RI*), devoid of repetitive sequences and which hybridized to the mouse probe *b*, was subcloned into pBR322. 900 bp from the *fin-2/H6RI* fragment and 270 bp from the murine probe *b* were sequenced according to the Sanger dideoxy method³⁰ using ³⁵S-labelling and gradient buffer sequencing gels³¹.

some 5 or of part of the long arm of this chromosome (deletion 5q⁻) has been reported in patients with refractory anaemia or acute non-lymphocytic leukaemia¹⁹. But other additional receptors or growth factors, including CSF-1, GM-CSF, IL-3 and PDGF-receptor, are clustered in the same region of chromosome 5 as *c-fms*^{20–22}. Consequently, it is not clear whether the haemopoietic disorders accompanying the 5q⁻ deletion are related to the loss of the *c-fms* and/or another growth factor or receptor gene¹⁷.

Another growth factor receptor, *c-erbB*, coding for the epidermal growth factor receptor, is activated during avian retrovirus-induced erythroleukaemias²³. Our study is the first to implicate a physiological haematopoietic growth-factor receptor in a leukaemogenic process. The myeloblastic tumours can be induced experimentally *in vivo* and *in vitro* and therefore represent a suitable system for analysing the contribution of the

CSF-1 receptor to the growth properties of the tumour cells. Although all the myeloblastic cell lines consist of non-adherent, factor-independent cells, of which more than 90% are blastic and myeloperoxidase-positive, only some of the myeloblastic tumours express significant amounts of *c-fms* mRNA. The *c-fms*-negative tumours therefore act as controls with which to compare the *c-fms*-positive tumours. If *c-fms* expression is important for the growth properties of the *c-fms*-positive tumours and the receptor is not mutated, it will be interesting to determine whether the cells stimulate their own growth in an autocrine fashion by synthesizing CSF-1 or a related molecule or respond using a paracrine mechanism with ligand synthesized in other cells.

We thank Paule Varlet and Martine Charon for technical assistance, Carine Pellet for typing the manuscript, Bernard Boursin for the illustrations, and Philip Avner, Frank Healy and Marc Sitbon for reviewing the manuscript and Douglas Lowy for helpful discussions. This work was supported by grants from the Centre National de la Recherche Scientifique and from the Association de la Recherche contre le Cancer.

Received 1 May; accepted 31 July 1987.

1. Nusse, R. *Trends Genet.* **2**, 244-247 (1986).
2. Sola, B., Fichelson, S., Bordereaux, D., Tambourin, P. E. & Gisselbrecht, S. *J. Virol.* **60**, 718-725 (1986).
3. Sherr, C. J. *et al. Cell* **41**, 665-676 (1985).

4. Heard, J. M. *et al. Molec. cell Biol.* **4**, 216-220 (1984).
5. Heard, J. M., Sola, B., Martial, M. A., Fichelson, S. & Gisselbrecht, S. *Blood* **68**, 193-199 (1986).
6. Peters, G., Lee, A. E. & Dickson, C. *Nature* **320**, 628-631 (1986).
7. Payne, G. S., Bishop, J. M. & Varmus, M. E. *Nature* **295**, 209-214 (1982).
8. Shen-Ong, G. L. C., Morse, H. C., Potter, M. & Mushinski, J. F. *Molec. cell Biol.* **6**, 380-392 (1986).
9. Weinstein, Y., Ihle, J. N., Laure, S. & Reddy, E. P. *Proc. natn. Acad. Sci. U.S.A.* **83**, 5010-5014 (1986).
10. Mushinski, J. F., Potter, M., Bauer, S. R. & Reddy, E. P. *Science* **220**, 795-798 (1983).
11. Müller, R. *et al. Molec. cell Biol.* **3**, 1062-1069 (1983).
12. McDonough, S. K., Larsen, S., Brodey, R. S., Stock, N. D. & Hardy, W. D. Jr *Cancer Res.* **31**, 953-956 (1971).
13. Wheeler, E. F. *et al. J. Virol.* **59**, 224-233 (1986).
14. Sacca, R., Standley, E. R., Sherr, C. J. & Rettenmeier, C. W. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3331-3335 (1986).
15. Sarma, P. S., Sharar, A. & McDonough, S. *Proc. Soc. exp. Biol. Med.* **140**, 1365-1368 (1972).
16. Groffen, J. *et al. Nucleic Acids Res.* **11**, 6331-6339 (1983).
17. Le Beau, M. M. *et al. Science* **231**, 984-987 (1986).
18. Pettenati, M. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 2970-2974 (1987).
19. Wisniewski, L. P. & Hirschhorn, K. *Am. J. Hemat.* **15**, 206-210 (1983).
20. Huebner, K. *et al. Science* **230**, 1282-1285 (1985).
21. Yang, Y.-C. *et al. Cell* **47**, 3-10 (1986).
22. Yarden, Y. C. *et al. Nature* **323**, 226-232 (1986).
23. Nilsen, T. W. *et al. Cell* **41**, 719-726 (1985).
24. Scherrer, K. in *Fundamental Techniques in Virology* (eds Habel, K. & Salzmann, N. P.) 413-432 (Academic, New York, 1969).
25. Donner, L., Fedele, L. A., Garon, C. F., Anderson, C. J. & Sherr, C. J. *J. Virol.* **41**, 489-500 (1982).
26. Fort, P. *et al. Nucleic Acids Res.* **13**, 1431-1442 (1985).
27. Roussel, M. F., Sherr, C. J., Barker, P. E. & Ruddle, F. H. *J. Virol.* **48**, 770-773 (1983).
28. Coussens, L. *et al. Nature* **320**, 277-280 (1986).
29. Eladari, M. E., Syed, S. H., Guilhot, S., d'Auriol, L. & Galibert, F. *Biochem. biophys. Res. Commun.* **140**, 313-319 (1986).
30. Sanger, F. S., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
31. Biggin, M. D., Gibson, T. S. & Hong, G. F. *Proc. natn. Acad. Sci. U.S.A.* **80**, 3963-3965 (1983).

Cytokine-induced pseudopodial protrusion is coupled to tumour cell migration

Raouf Guirguis, Inger Margulies, Giulia Taraboletti, Elliott Schiffmann & Lance Liotta

Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Pseudopodia protrusion is a prominent feature of actively motile cells *in vitro*^{1,2} and invading tumour cells *in vivo*³; however, the function and regulation of pseudopodia are poorly understood. Tumour autocrine motility factor (AMF) represents a new class of cytokines which are secreted by tumour cells⁴ and embryonic cells⁵ and induce random motility in the producer cells or in heterologous cells with appropriate receptors. Here we report that a major effect of this factor is to induce the extension of cell pseudopodia before cell translocation. Using a new method to quantify and isolate pseudopodia, we find that human breast carcinoma cell AMF (at concentrations of 1 nM or below) stimulates random pseudopodia formation in a dose-dependent and time-dependent manner. Anti-AMF antibodies inhibit pseudopodia protrusion and cell motility, showing the importance of pseudopodia formation during locomotion. AMF-stimulated motility and pseudopodia formation occur on a wide variety of adhesive substrata which suggests that certain intrinsic motility events are independent of the attachment mechanism. Induced pseudopodia show a prominent axial actin network in the electron microscope. The number of laminin receptor and fibronectin RGD recognition sites is increased by a factor of 20 in the induced pseudopodia when compared to the average distribution in unstimulated cells. Exploratory pseudopodia regulated by cell-derived motility factors contain receptors for matrix proteins and could serve as 'senseorgans' essential to the process of cell locomotion.

A class of cytokines has recently been described which are distinguished from growth factors by their ability to markedly stimulate random locomotion of cells⁴⁻⁶. They are heat-labile proteins of similar size and show some degree of cell-type specificity in both production and response. The autocrine motil-

Table 1 Effect of substratum on AMF-induced (1 mM) pseudopodia protrusion and whole cell migration

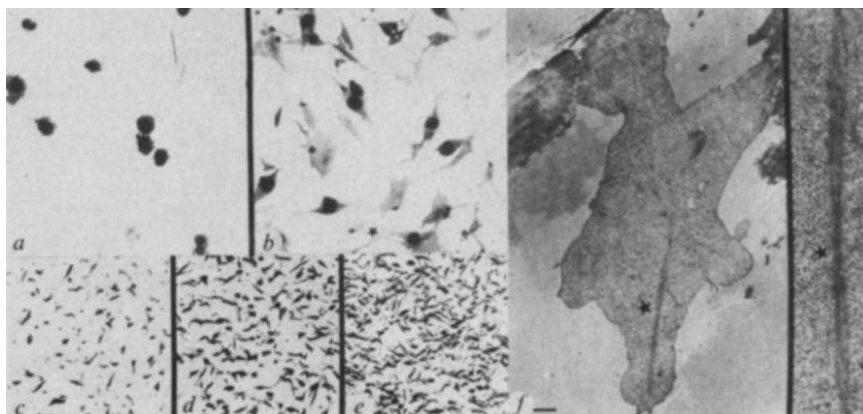
Substratum type (nucleopore filter coating)	AMF	Migrating cells ($\pm$ s.e.m. $n = 4$)	Induced pseudopodia (density units)
Uncoated polycarbonate	-	9 $\pm$ 2	0.09
Uncoated polycarbonate	+	75 $\pm$ 4	2.81
Laminin (10 μ g ml ⁻¹)	+	76 $\pm$ 3.7	2.64
Serum fibronectin (10 μ g ml ⁻¹)	+	72 $\pm$ 4.2	5.32
Type IV collagen (10 μ g ml ⁻¹)	+	90 $\pm$ 3.2	5.12
Type I collagen (10 μ g ml ⁻¹)	+	70 $\pm$ 2.8	3.12

Whole cell migration and pseudopodial protrusion was measured as described in Fig. 2 legend. Substratum coating was conducted as described previously⁴.

ity factor isolated from human melanoma cells⁴ or rat breast carcinoma cells⁶ is a protein of relative molecular mass (M_r) 54,000 (54K) which stimulates the random or directed migration of the producer cells and is augmented after *ras* family oncogene transfection⁴, but fails to act as a chemoattractant for granulocytes⁴. Response to AMF is blocked by treatment with pertussis toxin but not by agents which modulate or participate in the adenylate cyclase pathway⁷. Thus, cholera toxin and forskolin, which stimulate cyclic AMP formation, dideoxyadenosine, which inhibits adenylate cyclase, and 8-bromo-cAMP, which acts like cAMP, all have negligible effects upon motility⁷. The motility factor partially characterized from embryonic fibroblasts and some types of transformed fibroblasts⁵ (termed 'scatter factor') stimulates the random migration of epithelial cells, but fails to cause certain tumour cells to migrate, and has a molecular weight in the range of 50-60K.

Human breast carcinoma cell line MDA 435 produces a motility factor similar or identical to the AMF purified from human melanoma cells⁴. The breast carcinoma AMF markedly stimulates the random motility of the producer cells (Figs 1 and 2) and this is accompanied by a pronounced alteration in cell shape, including pseudopodia extension. To isolate induced pseudopodia, breast carcinoma cells in suspension in a modified

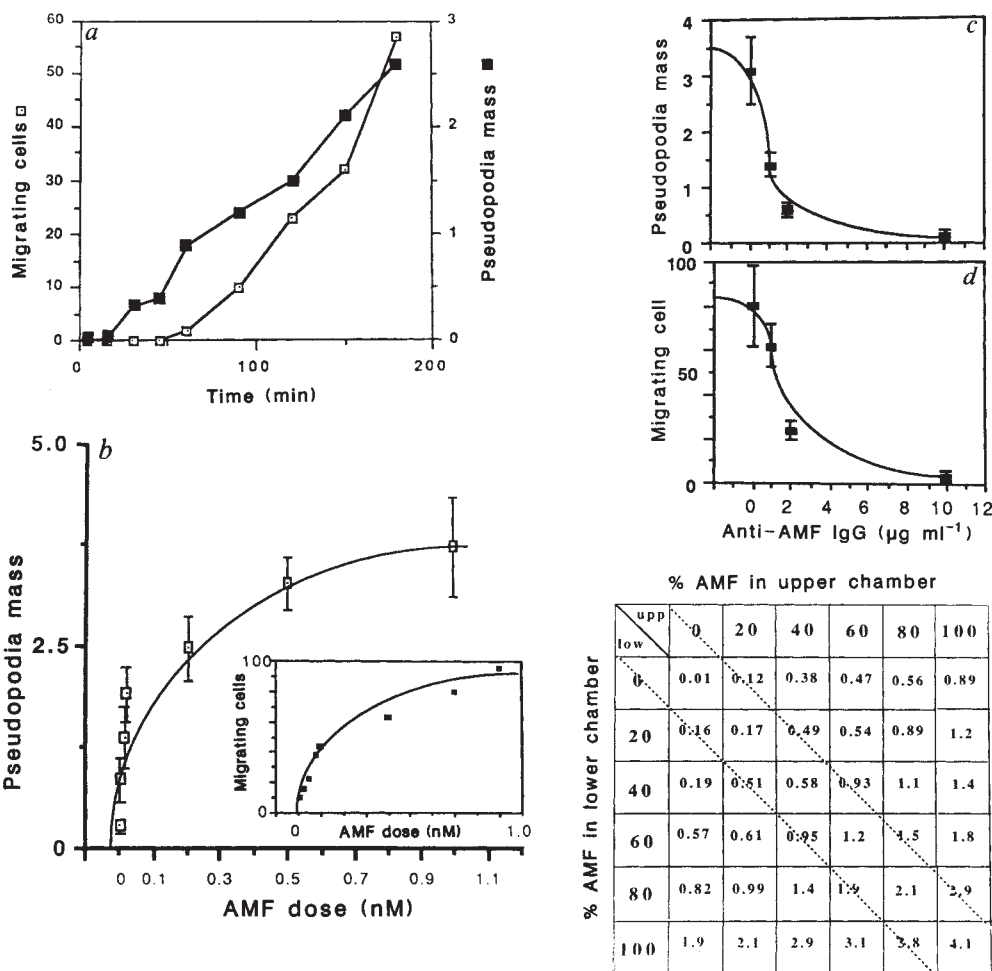
Fig. 1 Pseudopodia induction by autocrine motility factor (AMF) purified from the conditioned media of human breast carcinoma cells. *a*, MDA-435 human breast carcinoma cells, control (Dulbecco's minimal essential medium (DMEM)+10% fetal calf serum (FCS)) 5 h ($\times 440$ mag). The cells are attached and spread. *b*, MDA-435 cells treated with AMF (5 nM diluted in DMEM+10% FCS) for 5 h ($\times 440$ mag). Extensive alteration in cell shape with prominent pseudopodia (*) induction. *c*, *d*, *e*, Dose curve of AMF-induced pseudopodia protruding from the bottom of 3- μ m pore filters ($\times 44$). Pseudopodia were collected at 5 h, *c*, 0.2 nM AMF, *d*, 0.4 nM AMF, *e*, 2.0 nM AMF. *f*, Ultrastructure of AMF-induced protruding pseudopodia (bar, 1 μ m): an oblique section of the pseudopodia protruding through the filter (F). A prominent axial actin filament bundle (*) is shown enlarged in the insert.



Methods. AMF was purified from the serum-free conditioned media of MDA435 human breast carcinoma cells (ATCC no. HTB129) and assayed using a modified Boyden chamber migration assay using methods described previously for human melanoma cells⁴. The purified AMF migrated in SDS-PAGE as a single polypeptide chain of $M_r \sim 54K^4$. The N-terminal sequence of the purified AMF was DKELRFDRCTKSLAEANKK, with a repetitive yield of 92%. This sequence cannot be identified with any proteins in the N.I.H. amino-acid database. Pseudopodia protrusion was studied using Neuroprobe 48-well measuring whole cell migration⁴. MDA435 cells grown to 75% confluency were collected with trypsin-EDTA, allowed to recover at room temperature for 1 h in DMEM supplemented with 10% FCS and suspended at 2×10^6 cells ml^{-1} in DMEM with 1 mg ml^{-1} bovine serum albumin. Cells ($100 \mu l$) were added to the upper chamber and incubated at $37^\circ C$ for 5 h, with or without the indicated amounts of AMF. At the end of the incubation period, the filters (3 μ m pore Nucleopore polyvinylpyrrolidone-free) were removed from the chamber, stained with Diff-Quick, and transferred to a cleaned microscope slide, with the side of the filter containing the protruding pseudopodia apposed to the glass surface. After the cell bodies were wiped off the top of the filter, the pseudopodial protrusions adhered to the glass surface and could be counted in the microscope. For ultrastructural studies, the pseudopodia were fixed in 2.5% glutaraldehyde, embedded in epon, sectioned with an LKB microtome and examined using a Phillips 400 electron microscope operated at 60 kV.

Fig. 2 Random pseudopodial induction by AMF from tumour cells precedes whole cell translocation. *a*, Time course of pseudopodia protrusion compared to whole cell translocation. *b*, Dose dependence of both pseudopod protrusion and whole cell translocation (insert curve). *c*, 'Checkerboard' analysis of the appearance of tumour cell pseudopodia at 5 h in response to the specified AMF ratios in the upper and lower chamber (1 nM, 100%). Dose-dependent inhibition of pseudopodial protrusion and whole cell translocation by antibodies to AMF.

Methods. Tumour cell migration was measured as described previously⁴ using 8- μ m pore size Nucleopore filters in the chambers. Cell migration was quantified by recording the number of cells per high power field (400 $\times$) that had migrated through the filter after a 4 h period. Quantification of pseudopodia was achieved using an LKB laser densitometer with a model 220 Recording Integrator. Pseudopodial mass is shown in density units which exhibited a linear relationship with pseudopodia protein content (1.0 density unit, 120 μ g pseudopodial protein). In the 'Checkerboard' analysis the number of pseudopodia induced in a series of increasing uniform concentrations of AMF are shown in the diagonal. The lower triangle gives values for pseudopodia induced toward a positive gradient. Antibodies to AMF were prepared in New Zealand white rabbits with an initial subcutaneous immunization of 20 μ g purified AMF in complete Freund's complete adjuvant and boosted at subsequent 3-week intervals with 10 μ g. Bleeding was conducted 10 days after the third boost. Specificity of the antibodies were verified by immunoblotting and immunoprecipitation compared to preimmune sera. For the inhibition studies shown in *d*, the AMF was preincubated with the antibodies for 30 min before addition to the lower chamber.



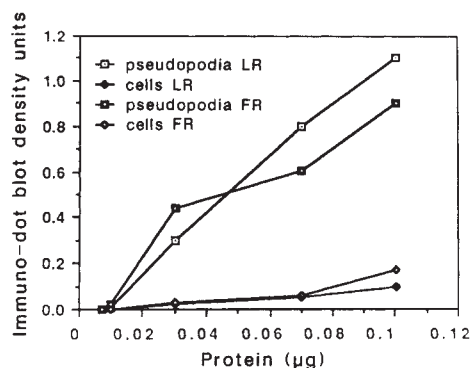


Fig. 3 Augmentation of laminin receptor (LR) and fibronectin receptor (FR) (RGD recognition) antigen on isolated pseudopodia. Receptor antigen content of detergent extracted isolated pseudopodial plasma membranes (pseudopodia) compared to whole cell plasma membranes (cell). Quantitative immuno-dot blotting was performed as described¹⁶. Each data point represents the mean of three determinations using matched 1/500 antibody dilutions and the indicated amount of extracted pseudopodia or cell protein. Anti-fibronectin receptor antibodies described previously^{8,9} were supplied by Dr W. T. Chen of Georgetown University, Washington, DC. Anti-laminin receptor monoclonal antibodies and plasma membrane extractions were prepared and characterized as described previously¹⁵.

Boyden chamber were attached to the surface of polycarbonate filters having a pore size which prevented whole cell traversal but allowed the protrusion of pseudopodia (Fig. 1c–e). The induced pseudopodia were collected from the opposite face of the filter by adherence to glass (Fig. 2) or by mechanical scraping. This method demonstrated that pseudopodia protrusion was induced by AMF in a time- and dose-dependent manner (Fig. 2) and began ~30 min before whole cell translocation. Preincubation of AMF with anti-AMF antibodies inhibited the induction of pseudopodia in parallel with inhibition of motility (Fig. 2d).

Checkerboard analysis (Fig. 2c) was used to establish whether AMF induced random or directed pseudopodia formation. When AMF was introduced into either or both chambers, pseudopodia protrusion was extensive, so we conclude that AMF stimulates random pseudopodia formation and does not require a gradient of AMF to orient the pseudopodia. This agrees with the finding that AMF also stimulates random motility to a much greater degree than directed motility⁴.

Ultrastructural studies of pseudopodia induced by AMF showed that they frequently contained bundles of actin cables along the long axis of the pseudopod (Fig. 1). The actin cables terminated at the inner membrane zone of the leading edge of the pseudopodia. It has been proposed that cytoplasmic actin bundles bind indirectly to the cytoplasmic domain of transmembrane extracellular matrix (ECM) receptors^{8–10}. Furthermore, ultrastructural studies of tumour cells migrating through the rat mesentery have suggested that the tips of cell pseudopodia form receptor-mediated contacts with the ECM³. AMF-stimulated pseudopodia are an ideal system to study whether or not ECM receptors pre-exist on the pseudopodia or are induced only when the pseudopod is adherent to the ECM component. It has been suggested that ECM receptors are randomly distributed on the surface of the cell and that the anchorage of the receptor to the extracellular ligand regulates the distribution of the receptor. But in the migrating tumour cell, which involves reversible attachment as part of the locomotory response, the distribution of ECM receptors is not random. When AMF stimulates the formation of pseudopodia before the locomotion cascade, an asymmetric distribution of receptors occurs, with a higher concentration in the pseudopods than in the membranes of unstimu-

lated cells. We find that AMF-induced isolated pseudopodia contain 20 times more laminin and fibronectin Arg-Gly-Asp (RGD) recognition receptors (Fig. 3) than the average number found in the plasma membrane of unstimulated cells. AMF can stimulate pseudopodia formation and migration on a variety of adhesive substrata, including uncoated polycarbonate filters or polycarbonate filters coated with fibronectin, laminin, type IV collagen, or type I collagen (Table 1). As pseudopodia can be induced by AMF in the absence of a laminin- or fibronectin-coated substratum, the increase in the number of receptors on the pseudopodia must precede their interaction with extrinsic laminin and fibronectin.

Pseudopodia induction by soluble factors has been previously limited to polymorphonuclear leukocytes migrating in a gradient or uniform concentrations of a chemoattractant¹¹. This study demonstrates that uniform concentrations of a cytokine are sufficient to induce pseudopodia formation and subsequent translocation of the cell. We conclude that pseudopodia formation is an integral component of cytokine-regulated tumour cell locomotion. Pseudopodia regulation can occur in a manner independent of the adhesion mechanism, although some type of adhesion is required for cell traction^{12,13}. Autonomous random cell motility could begin with the induction of exploratory pseudopodia before cell translocation. ECM receptors concentrated on the surface of the pseudopodia might sense the distribution of ECM components in the immediate cellular environment and in turn help determine the vectorial direction of locomotion.

Received 9 July; accepted 6 August 1987.

- Couchman, J. R., Lenn, M. *Eur. J. Cell Biol.* **36**, 182–194 (1985).
- Mohler, J. L., Partin, A. W., Isaacs, W. B. & Coffey, D. S. *J. Urol.* **137**, 544–547 (1987).
- Muller, G. W., Haemmerli, G. & Strauli, P. *Cell Biol. Int. Rep.* **9**, 447–461 (1985).
- Liotta, L. A. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 3302–3306 (1986).
- Stoker, M., Gherardi, E., Perryman, M. & Gray, J. *Nature* **327**, 239–242 (1987).
- Atnip, K. D., Haney, L., Nicolson, G. L. & Dabbous, M. K. *Biochem. biophys. Res. Commun.* **146**, 996–1002 (1987).
- Stracke, M. L., Guirguis, R., Liotta, L. A. & Schiffmann, E. *Biochem. biophys. Res. Commun.* **146**, 339–345 (1987).
- Chen, W. T., Wang, J., Hasegawa, T., Yamada, S. S. & Yamada, K. M. *J. Cell Biol.* **103**, 1649–1661 (1986).
- Chen, W. T., Chen, J. M. & Mueller, S. C. *J. Cell Biol.* **103**, 1073–1090 (1986).
- Ruoslahti, E. & Pierschbacher, M. D. *Cell* **44**, 517–518 (1986).
- Gallin, J. I., Gallin, E. K., Malech, H. L. & Cramer, E. B. in *Leukocyte Chemotaxis* (eds Gallin, J. I. & Quie, P. G.) 123–141 (Raven, New York, 1978).
- Pryzwansky, K. B., Schliwa, M. & Porter, K. R. *Eur. J. Cell Biol.* **30**, 544–547 (1983).
- Trinkaus, J. P. *J. Neurosci. Res.* **13**, 1–19 (1985).
- Carter, S. B. *Nature* **208**, 1183–1187 (1965).
- Wewer, U. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 7137–7141 (1986).
- Towbin, H. & Gordon, J. *J. immun. Meth.* **72**, 313–340 (1984).

Importance of DNA stiffness in protein–DNA binding specificity

M. E. Hogan* & R. H. Austin†

* Department of Molecular Biology and † Department of Physics, Princeton University, Princeton, New Jersey 08544, USA

From the first high-resolution structure of a repressor bound specifically to its DNA recognition sequence¹ it has been shown that the phage 434 repressor protein binds as a dimer to the helix. Tight, local interactions are made at the ends of the binding site, causing the central four base pairs (bp) to become bent and overtwisted. The centre of the operator is not in contact with protein but repressor binding affinity can be reduced at least 50-fold in response to a sequence change there². This observation might be explained should the structure of the intervening DNA segment vary with its sequence, or if DNA at the centre of the operator resists the torsional and bending deformation necessary for complex formation in a sequence dependent fashion. We have considered the second hypothesis by demonstrating that DNA stiffness

is sequence dependent. A method is formulated for calculating the stiffness of any particular DNA sequence, and we show that this predicted relationship between sequence and stiffness can explain the repressor binding data in a quantitative manner. We propose that the elastic properties of DNA may be of general importance to an understanding of protein-DNA binding specificity.

The fact that DNA behaves as a stiff elastic rod has been recognized since Peterlin's early analysis of light scattering data³. This rigidity is expressed as the persistence length P of the molecule, which is related to the Young's modulus, E (ref. 4):

$$P = EI/kT \quad (1)$$

where I is the surface moment of inertia for a right cylinder, k is Boltzmann's constant and T is the absolute temperature; E is related to the stress which develops when the long axis of a rod is strained. The intrinsic shear modulus, G , relates the restoring stress when a lateral or torsional strain is applied to the material. The Young's and shear moduli are related by Poisson's ratio μ , which is also intrinsic:

$$G = E/2(1 + \mu) \quad (2)$$

where μ is close to 0.5 for polymeric materials^{5,6}.

Recently, it has been found that E and G are not constant for DNA, but are a function of bp composition⁷⁻¹⁰ which is not surprising as the stacking and the hydrogen bonding stabilizing DNA are strongest for G-C bps. The available experiments do not measure the intrinsic moduli directly, but instead the extrinsic bending constant B and the torsional constant C , given by;

$$C = IG/L \quad (3)$$

and

$$B = EIL \quad (4)$$

where L is the length of the molecule under strain and I is the area moment of inertia.

In Table 1 we present measured values for the Young's modulus of DNA as a function of base composition. Although the modulus E is the fundamental parameter, physical techniques measure EI . To extract E from those data, we have calculated I , using the crystallographic radius for B-DNA (11 Å), although it will be shown that the significant conclusions to be drawn from these analyses are not dependent on the value chosen. It is clear that the available data are not comprehensive, and the theories used to calculate EI from the data not perfect, especially in analysis of anisotropy and quenching data. Consequently, when the individual methods are applied to similar random sequence DNA samples, calculated values for E can vary by as much as a factor of three. Nevertheless, there is a good qualitative agreement among the various methods which

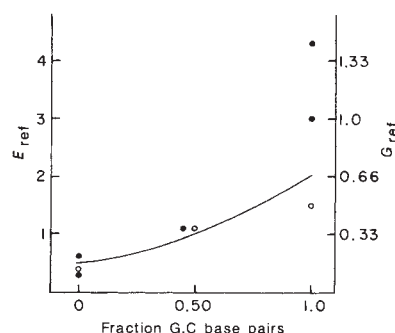


Fig. 1 Summary of the relation between stiffness and DNA sequence. The normalized Young's moduli E catalogued in Table 1 have been plotted against the mole fraction of G-C bps in the helix being studied. The corresponding shear moduli (G_{ref}) have been calculated from the Young's modulus, as described in the text. Solid line, compositional dependence predicted by equations (9) and (10) for random DNA sequences, using the consensus stiffness parameters in Table 2. ●, Measurements on random sequences or non-alternating synthetic helices, such as poly(dA).poly(dT); ○, measurements on DNA sequences with a simple alternating sequence, such as poly(dA-dT).poly(dA-dT) or poly(dG-dC).poly(dG-dC).

we have emphasized by normalizing the outcome of each, so that for a random sequence DNA (45% G + C) the Young's modulus is adjusted to a traditional value (1.1×10^9 dyn cm⁻²), corresponding to a persistence length of 175 bps at 300 K. Although this normalization is not strictly valid, the arguments presented below show this first order correction provides a useful insight into the properties of DNA.

The shear modulus of DNA, G , is also presented in Table I, calculated from the Young's modulus using equation (2) and assuming that μ is 0.5, which has previously given close agreement with G as measured for DNA⁷. The relation does not hold for the homopolymer poly(dA).poly(dT) (ref. 7), but this discrepancy could result from the atypical B-helix structure assumed by poly(A) sequences¹¹. Because of this ambiguity, and because the Young's modulus of DNA has been measured by several methods whereas the shear modulus has not, our opinion is that, at present, the sequence dependence of DNA stiffness is best evaluated from calculations using the Young's modulus (as measured by light scattering, pulsed birefringence, triplet anisotropy decay and triplet quenching).

A conservative interpretation of the available data suggests that, at the extremes of base composition, E , and hence the bending and torsional stiffness of DNA, can vary by at least a factor of four (Fig. 1; Table 1).

Table 1 DNA flexibility measurements

Method Sample	Triplet anisotropy $E(E_{ref})$	Birefringence $E(E_{ref})$	Light scatter $E(E_{ref})$	Triplet quenching $E(E_{ref})$	Consensus values E G	
poly(dG).poly(dC)	3.0 (2.4)			1.33 (4.3)	2	0.66
poly(dG-dC).poly(dG-dC)			1.5 (1.5)		2	0.66
poly(dA-dC).poly(dG-dT)	1.4 (1.1)				1	0.33
Random	1.4 (1.1)	1.1 (1.1)	1.1 (1.1)	0.34 (1.1)	1	0.33
poly(dA).poly(dT)	0.8 (0.6)			0.10 (0.32)	0.5	0.167
poly(dA-dT).poly(dA-dT)		0.36 (0.36)			0.5	0.167

These data are a compilation of data available in the literature. In each instance the Young's modulus E or shear modulus G has been presented in units of 10^9 dyne cm⁻². For each method which has been listed, a random sequence DNA was also measured. To compare the sequence dependence of those data more directly, we have normalized measured values E such that for each method, random sequence DNA is assigned the value 1.1×10^9 . Those normalized moduli are presented in the parenthesis as E_{ref} . Triplet anisotropy⁷ pulsed electric birefringence decay⁸ light scattering⁸ and triplet lifetime quenching¹⁰ measurements are cited. For the birefringence and light scattering data, the original authors have calculated a persistence length for DNA. That value has been converted to a Young's modulus using equation (1) in the text and $I = 2.3 \times 10^{-28}$ cm⁴. Consensus values for the Young's moduli E are an evaluation of the available literature and may be refined in the context of future experiments. Consensus values for G have been calculated using equation (2).

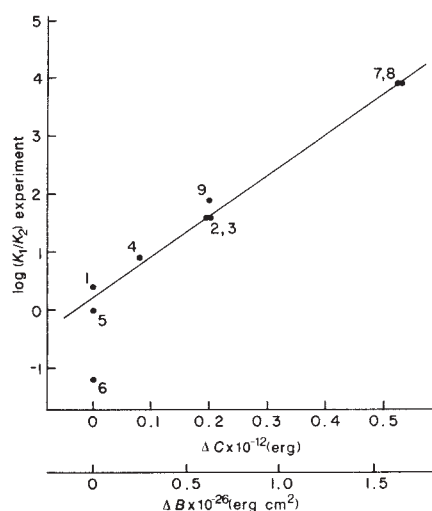


Fig. 2 Sequence dependence of 434 repressor binding: theory compared with experiment. The measured change in repressor binding affinity (from ref. 2) is compared to the change in DNA stiffness calculated from theory. See Table 2 for the numerical values. Experimental data are presented as $\log(K_1/K_2)$ plotted against ΔC or ΔB . Individual pairs (data against theory for each operator sequence) are identified by the numbering system in Table 2. Because C and B are proportional, their correlation with experiment has been presented together, by appropriate labelling of the abscissa. Solid line, linear least-squares analysis, assuming that the error in $\log(K_1/K_2)$ is ± 0.2 , as ascertained by the scatter of the binding data. A good linear correlation of that kind suggests that the measured binding constant change is determined by DNA stiffness effects.

Ethidium bromide fluorescence anisotropy experiments have surprisingly failed to detect sequence-dependent effects on the shear modulus^{6,12}. The discrepancy between the ethidium bromide data and other physical measurements of DNA stiffness is a source of continuing controversy, but could be related to the fact that short-lived fluorescence decay measurements are most sensitive to high-frequency (local) motions of DNA, and for that reason are susceptible to changes in the DNA molecule which occur near the dye probe binding site. High-frequency DNA motions may be crucial to an understanding of its stability and function, but we will limit our present discussion of stiffness variation to results verified by several techniques.

To understand protein binding specificity in terms of DNA stiffness variation, it is first necessary to consider the work required to deform a helix. The mechanics of elastic media specifies that if a length L of DNA, with Young's modulus E and area moment of inertia I is strained by bending into an arc of radius S or twisting through an angle ϕ (in radians) the work done is;

$$U(\phi) = -IG\phi^2/2L = -C\phi^2/2 \quad (5)$$

and

$$U(S) = -EIL/2S^2 = -B/2S^2 \quad (6)$$

This linear elastic model may be too simple, because it overlooks the possibility that DNA bending and twisting are directional¹³ and assumes that DNA can be modelled as a continuum rod. But, given that the same simplification is built into theories used to interpret the hydrodynamic and scattering data (see Table 1), the above equations are adequate approximations.

To complete our model, it is necessary to describe how the contribution of individual bps sums to produce a net bending constant B and torsional constant C . DNA flexibility as measured experimentally does not involve a disruption of bps but is a property of the duplex state, so sequence-dependent stiffness variation must be a function for the vertical (base-

stacking) interaction between the bps. To a first approximation then, the stiffness of a DNA segment should vary as the frequency of nearest neighbours in the helix. Thus, a helix can be modelled as a heterogeneous elastic rod with E and G varying with the type of nearest neighbours. Equilibrium statics says that for a heterogeneous rod composed of N equally spaced subunits, the net torsional constant is:

$$1/C = 1/C_1 + 1/C_2 + \dots 1/C_n \quad (7)$$

and the bending constant is:

$$1/B = 1/N^2(1/B_1 + 1/B_2 + \dots 1/B_n) \quad (8)$$

where the subscript n refers to the individual elements of the rod. If, as the data suggest (Fig. 1), DNA stiffness is dominated by the nearest-neighbour interactions and is relatively insensitive to bp inversion we can usefully write the above equations;

$$1/C = \frac{L}{I} (f_{aa}/G_{aa} + f_{ag}/G_{ag} + f_{gg}/G_{gg}) \quad (9)$$

and

$$1/B = \frac{1}{IL} (f_{aa}/E_{aa} + f_{ag}/E_{ag} + f_{gg}/E_{gg}) \quad (10)$$

where L is the overall length of the helix segment in cm, G_{aa} (for example) is the shear modulus between two adjacent A·T bps and f_{aa} is the frequency of AT nearest neighbours, normalized to unity. The value for each element of E or G has been assigned by reference to the consensus values listed in Table 1. Thus, for any DNA segment

$$f_{aa} + f_{ag} + f_{gg} = 1 \quad (11)$$

Using the above formalism, we can evaluate the contribution of DNA stiffness variation to the binding specificity of the 434 repressor. In Table 2 we have catalogued the three base-stacking interactions at the centre of the operator for each of the synthetic binding sites studied by Koudelka *et al.*² Based upon those nearest neighbour frequencies, we have used equation (9) to calculate the bending and torsional spring constant for the region (Table 2). Next, as a first approximation, we propose that when DNA sequence is changed at the centre of the operator (where there are no protein contacts), the resulting change in repressor affinity is due exclusively to a change in the binding free energy which must be spent to bend or twist the central DNA segment. This free energy is stored as bending and torsional strain, as described by equations (5) and (6). If two particular operator sequences labelled 1 and 2 have torsional strain energies $U_1(\phi)$ and $U_2(\phi)$ stored for a fixed angular strain ϕ , then the dissociation constant change which results from the stiffness difference can be expressed as the ratio K_1/K_2 :

$$K_1/K_2 = \exp[2(U_2(\phi) - U_1(\phi))/kT] \quad (12)$$

Further, by taking the log of both sides of the above equation and substituting in equation (5) we find:

$$\log(\{K_1/K_2\}) = \phi^2/kT \times (C_2 - C_1) \quad (13)$$

Similarly, in terms of bending energetics:

$$\log(K_1/K_2) = 1/kTS^2 \times (B_2 - B_1) \quad (14)$$

Note that in equations (13) and (14) we assume that the twist angle ϕ and the bending radius S each represent one degree of freedom and that the energy of mechanical deformation has a negligible entropy contribution.

The calculated torsion and bending constants are presented for each synthetic operator in Table 2, along with the binding data measured by Koudelka *et al.*² The correlation between theory and experiment is presented in Fig. 2, which shows that the stiffness model accounts well for the sequence dependence of the data. It is interesting that the one sequence permutation which is not adequately fitted by these stiffness calculations is

Table 2 A calculated relation between 434 repressor binding and stiffness

Operator number	Name	Central sequence	f_{aa}	f_{ag}	f_{gg}	$C \times 10^{-12}$	$B \times 10^{-26}$	K_1/K_2 exper	$\log\{K_1/K_2\}$ exper
0	14	ATAT	1	0	0	0.37	1.15	1	0
1	6T	TTAA	1	0	0	0.37	1.15	1.5	0.4
2	6C	CTAG	1/3	2/3	0	0.57	1.8	5	1.6
3	6G	GTAC	1/3	2/3	0	0.57	1.8	5	1.6
4	6G/2	GTAT	2/3	1/3	0	0.45	1.4	2.5	0.9
5	7A	AATT	1	0	0	0.37	1.15	1	0
6	7A/2	AAAT	1	0	0	0.37	1.15	0.3	-1.2
7	7C	ACGT	0	2/3	1/3	0.9	2.8	50	3.9
8	7G	AGCT	0	2/3	1/3	0.9	2.8	50	3.9
9	7G/2	AGAT	1/3	2/3	0	0.57	1.8	7	1.9

The binding data of Koudelka *et al.*² have been summarized here. The central DNA sequence of each 14-bp long operator has been specified (positions 6–9). All other nucleotides are unchanged relative to operator 14, which is: ACAATATATATTGT. Binding data (K_1/K_2 experimental) are presented as the repressor dissociation constant, relative to that measured with the unmodified binding site (fragment 14). For example, 50 signifies that in ref. 2, 434 repressor binding to operator 7G was found to be 50 times weaker than to unmodified operator fragment 14. f_{aa} , f_{ag} , f_{gg} refer to the frequency of nearest neighbour associations at the centre of each synthetic operator. C is the calculated torsional spring constant for that central region in units of 10^{-12} dyne, derived by assuming $G_{aa} = 0.167 \times 10^9$ dyne cm^{-2} ; $G_{ag} = 0.33 \times 10^9$ dyne cm^{-2} ; $G_{gg} = 0.66 \times 10^9$ dyne cm^{-2} . B is the calculated bending constant for the central region, in units of 10^{-26} dyne cm^2 , derived by assuming $E_{aa} = 0.5 \times 10^9$ dyne cm^{-2} ; $E_{ag} = 1.0 \times 10^9$ dyne cm^{-2} ; $E_{gg} = 2.0 \times 10^9$ dyne cm^{-2} . See text for details.

7A/2 (number 6 in Fig. 2). That manipulation creates a three-nucleotide oligo-(A) sequence at the centre of the synthetic operator (Table 2). Such oligo-(A) segments have been described as the source of permanent helix bends^{11,14,15}. Our simple elastic analysis has probably failed to account for such a secondary structural effect in operator 6.

Although Anderson *et al.* point out that the centre 4 bp of the 434 repressor complex is substantially bent and overtwisted¹, the absolute change in twist or curvature cannot yet be determined as the structure of the uncomplexed binding site has not been measured. In the absence of this information, it is useful to quantify the relationship between theory and experiment in Fig. 2 in the context of two limiting models. If it is presumed that the sequence dependence of repressor binding in ref. 2 is determined only by torsional stiffness differences, the slope of Fig. 2 specifies that, within the central 4 bp, 32° of twist change will account for all the measured affinity change among the 10 synthetic binding sites. Also, in terms of bending, a 33 Å radius of curvature would account for all the measured binding constant difference (excluding fragment 6).

From the Anderson *et al.* data¹ it is clear that the bend and twist change occurring on binding 434 repressor is about half that specified by the two limiting models. Consequently, when bending and torsional flexibility are both considered, it is likely that the simple stiffness formalism which we have presented can account for the sequence dependence of 434 repressor binding to its operator sequence, both qualitatively and quantitatively.

Additional experimentation is required to understand the rules relating DNA sequence to DNA flexibility, but the data at hand suggest that the sequence dependence of DNA stiffness may be significant whenever a protein (or other ligand) must bend or twist DNA to form its bound complex. Within the limits of the model, the formalism which we have presented can be used to assess the contribution of bending or twisting energetics to this class of site-specific binding interaction.

This work benefited from discussion with Professor Gerry Manning and Professor Robert Hopkins, who has independently developed a similar model for DNA rigidity as a function of bp composition. R.H.A. acknowledges support from the Office of Naval Research, and M.E.H. from the NSF and the National Cancer Institute.

- Barkley, M. D. & Zimm, B. H. *J. chem. Phys.* **70**, 2991–3007 (1979).
- Millar, D. P., Robbins, R. J. & Zuwall, A. H. *J. chem. Phys.* **76**, 2080–2094 (1982).
- Hogan, M., LeGrange, J. & Austin, B. *Nature* **304**, 752–754 (1983).
- Thomas, T. J. & Bloomfield, V. A. *Nucleic Acids Res.* **11**, 1919–1931 (1983).
- Chen, H. H., Rau, D. C. & Charney, E. *J. biomolec. struct. Dynam.* **2**, 709–719 (1985).
- Berkoff, B., Hogan, M., LeGrange, J. & Austin, R. *Biopolymers* **25**, 307–316 (1986).
- Wu, H. M. & Crothers, D. M. *Nature* **308**, 509–513 (1984).
- Fujimoto, B. S., Shibata, J. H., Schurr, R. L. & Schurr, J. M. *Biopolymers* **24**, 1009–1022 (1985).
- Fratini, A. V., Kopka, M. L., Drew, H. R. & Dickerson, R. E. *J. biol. Chem.* **257**, 14686–14707 (1982).
- Ulanovsky, L., Bodner, M., Trifonov, E. N. & Choder, M. *Proc. natn. Acad. Sci. U.S.A.* **83**, 862–866 (1986).
- Hagermann, P. *Nature* **321**, 449–450 (1986).

Proline isomerism in staphylococcal nuclease characterized by NMR and site-directed mutagenesis

Philip A. Evans*§, Christopher M. Dobson*,
Roger A. Kautz†§, Graham Hatfull‡ & Robert O. Fox†

* Inorganic Chemistry Laboratory, University of Oxford,
South Parks Road, Oxford OX1 3QR, UK

† Department of Cell Biology, Stanford University Medical School,
Stanford, California 94305, USA

‡ Department of Molecular Biophysics and Biochemistry,
Yale University, New Haven, Connecticut 06520, USA

Nuclear magnetic resonance (NMR) studies have shown that two distinct folded conformations of staphylococcal nuclease coexist in solution¹ and that these two states can interconvert directly without passing through an unfolded state. These experiments have also revealed that the two forms have very different folding kinetics, although the possibility that one component is an obligatory intermediate for the folding of the other form could be discounted¹. Here we report NMR data which show that alternative unfolded states are also distinguishable. These observations led us to hypothesize that *cis/trans* isomerism at a single peptide bond between a proline and its preceding residue might be the origin of the conformational multiplicity. Proline 117 was identified as a likely candidate for the site concerned and a mutant protein, in which Pro 117 was replaced by Gly, was constructed in order to test this. Alternative conformations are not observed in the spectrum of this mutant, lending powerful support to this hypothesis.

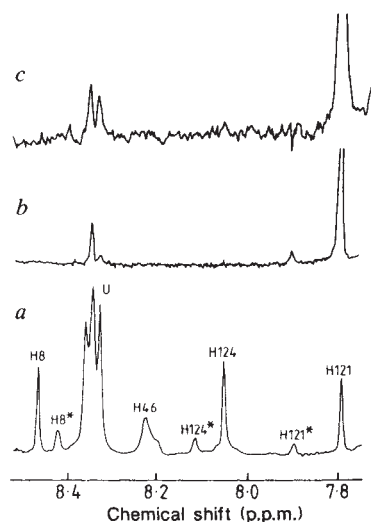
The existence of alternative folded states of nuclease is

Received 15 June; accepted 3 August 1987.

- Anderson, J. E., Ptashne, M. & Harrison, S. C. *Nature* **326**, 846–852 (1987).
- Koudelka, G. B., Harrison, S. B. & Ptashne, M. *Nature* **326**, 886–888 (1987).
- Peterlin, A. *Nature* **171**, 259–262 (1953).
- Landau, L. & Lifshitz, E. M. in *Statistical Physics* (Addison-Wesley, Reading, 1958).

§ Present addresses: Medical Molecular Biology Unit, Middlesex Hospital Medical School, Cleveland Street, London W1P 6DB, UK (P.A.E.) and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA (R.A.K.).

Fig. 1 *a*, Low-field region of the 500-MHz NMR spectrum of nuclease at pH 5.3, 48 °C in D₂O recorded as described previously¹. The assignments of the histidine resonances in the major and minor (*) states of the folded protein are indicated with H8 and H8* for His 8, and so on; the resonances between 8.3 and 8.4 p.p.m. arise from the histidine residues in the unfolded state. *b*, Differences between *a* and the spectrum resulting from presaturation of H121 for 2.5 s before the observation pulse. *c*, Difference between the spectrum recorded with H121 presaturated for 2.5 s and that recorded with H121 and H121* simultaneously presaturated for 2.5 s. The vertical scales of *b* and *c* are expanded by factors of 4 and 16 respectively to show the resonances more clearly.



revealed most clearly by the observation that in the ¹H NMR spectrum of the protein there are a number of minor resonances in addition to those attributed to the major folded form¹. These are particularly clearly observed for the H^ε resonances of the four histidines (Fig. 1*a*). These resonances have now been specifically assigned by comparing spectra of nuclease mutants in which individual histidines had been replaced by other residues (R.A.K. and R.O.F., in preparation; J. L. Markley, personal communication). The chemical shifts and the pH titration curves of the major and minor histidine resonances are closely similar suggesting that the major and minor forms of the protein do not have major differences in structure¹.

Our approach to investigating the relationships between the different conformational states of nuclease has been to probe their processes of interconversion, using magnetization transfer (MT) techniques. The basis of these experiments has been described previously^{1,2}. Figure 1 shows the effects of saturation of the resonances of His 121 H^ε at a temperature close to the midpoint of the thermal unfolding transition. Saturation of the major component alone (Fig. 1*b*) causes a diminution in the intensity of the minor resonance, reflecting interconversion of the two forms. In addition there are effects on two resonances amongst the cluster which arises from the His H^ε protons of the unfolded protein. This result demonstrates that His 121 can exist in two distinct environments not only in folded states but also when the protein is unfolded. The absence of broadening of the resonances shows that interconversion of these unfolded forms must be slower than ~1 s⁻¹. In a double saturation experiment¹, shown in Fig. 1*c*, in which the minor His 121 resonance is already saturated, irradiation of the major one is again observed to perturb the resonances of both unfolded states. This confirms that both these states are in exchange with the major folded form under these conditions. Similar heterogeneity in the unfolded state was not observed for the other histidines, suggesting that the disparity is more localised than that between the folded states of the nuclease.

This observation of distinguishable unfolded forms provides a crucial clue to the origin of the multiplicity of conformational states. Although it is possible to conceive of various possibilities for slowly interconverting alternative folded structures, much less scope exists when considering the unfolded population. Although the unfolded protein may not behave quite as a random coil, it certainly comprises a large number of rapidly interconverting conformations. It is highly unlikely that it could retain

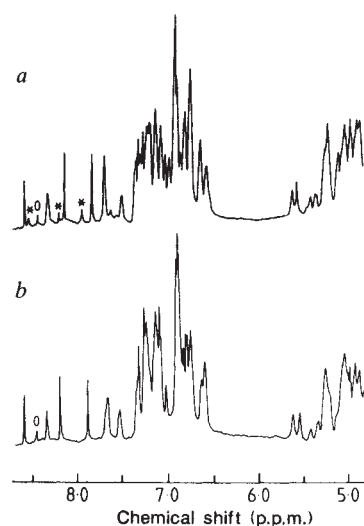


Fig. 2 The 500-MHz NMR spectra at pH 5.3, 40 °C in D₂O of *a*, wild-type nuclease and *b*, mutant nuclease with Pro 117 replaced by Gly 117. The resonance labelled 0 arises from formate impurity in the sample.

sufficient folded structure to give rise to distinct, slowly exchanging states. On the other hand, one phenomenon known to give rise to multiple unfolded forms of proteins is *cis-trans* isomerism of X-proline peptide bonds (where X is any residue)^{3,4}. There are six prolines in the nuclease molecule but we were able to pinpoint Pro 117 as a residue for particular consideration. This is because detailed investigation of the exchange kinetics suggests that in the major conformation of the folded protein it is the *cis* isomer which is present; in the crystal structure of the nuclease-Ca²⁺-thymidine 3',5'-disphosphate complex, only Pro 117 is observed to be in the *cis* orientation⁶. This evidence will be discussed fully elsewhere.

The probable involvement of Pro 117 is further supported by the fact that this is the proline residue closest in the sequence to His 121, the only histidine residue to be spectroscopically sensitive to the isomerism in the unfolded protein, and the most sensitive to isomerism in the folded state. To test the hypothesis further, however, the effects of the substitution of proline by a different residue at this position were investigated. In small peptides, at least, proline is unique amongst the common amino acids in its tendency to form *cis* and *trans* oriented peptide bonds of comparable stability^{3,4}. The equilibrium for peptide bonds which do not have proline at the second position strongly favours the *trans* isomer. It was therefore anticipated that if the model were correct, a significant shift in the conformational equilibrium, at the very least, ought to result from such a mutation.

The nuclease mutant, Pro 117→Gly, was obtained through oligonucleotide directed mutagenesis (R.A.K. and R.O.F., in preparation). This protein was found to possess nuclease activity, suggesting that it folds in a similar manner to the wild-type. This was further supported by the ¹H NMR spectrum which is shown in comparison to that of the wild-type in Fig. 2. The broad similarity of the spectra is evident in, for example, the chemical shifts of the major His H^ε resonances and in the characteristically shifted α-CH resonances between 5 and 6 p.p.m. This is powerful evidence for similar structures of the two proteins. Study of the temperature dependence of the NMR spectrum reveals that the thermal stability of the folded state was actually increased by about 5 °C as a result of the mutation.

The striking difference which is apparent in the spectrum of the mutant is the absence of the minor resonances which characterize that of the wild-type protein. This shows that whereas the mutation has had only a limited effect in the general fold of the

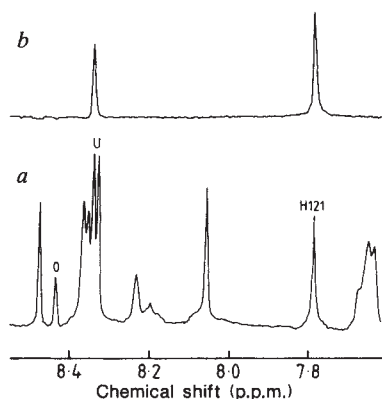


Fig. 3 *a*, Spectrum of the Pro 117 → Gly mutant at pH 5.3, 52 °C in D₂O and *b*, difference between this and the spectrum resulting from saturation of H121 for 2 s before accumulation.

molecule, the equilibrium observed in the wild-type protein between two variants of that fold has been strongly perturbed. Specifically, the absence of observable minor resonances must mean either that a single form is entirely predominant in the case of the mutant or, possibly, that interconversion between alternative forms is much faster than in the wild-type, so that only a single 'average' resonance is detected. At any rate, the pronounced effect of this sequence change, the like of which was not observed in the spectra of other mutants such as those used in assigning the histidine resonances (R.A.K. and R.O.F., in preparation), strongly supports the hypothesis that the conformational heterogeneity of the wild type is associated with isomerism at Pro 117.

This interpretation was further consolidated by study of the unfolded state of the mutant nuclease. Figure 3 shows the effect of saturation of the His 121 H^{ε1} resonance, close to the midpoint of thermal unfolding. It is clear that only a single resonance of this proton in the unfolded state is perturbed through an MT effect, which contrasts strongly with the analogous experiment on the wild-type, in which two such resonances were perturbed (Fig. 1). In the unfolded state as well, therefore, replacement of Pro 117 by Gly prevents the co-existence of alternative, slowly interconverting forms. This is precisely what would be anticipated, given the unique properties of X-proline linkages in small peptides³.

The significance of proline isomerism for staphylococcal nuclease, not only in the dynamics of folding but in the native state itself, is thus apparent. In the crystal structure of the nuclease-inhibitor complex Pro 117 is *cis*, and is located in a type-VI reverse turn⁵. Our experiments suggest that this conformation is similar to that of the major form in the absence of inhibitor. In the minor folded form where the proline is *trans* the turn conformation must clearly be different but the sensitivity of His-8, -121 and -124 as well as other unidentified resonances to the disparity between the two forms shows that small changes are, in fact, propagated extensively through the structure as a result of isomerization. In the unfolded state, the sensitivity of His 121 to the isomerism of the bond between residues 116 and 117, though much smaller than in the folded state, is itself remarkable because these residues are by no means constrained by the sequence to be close in space in the absence of folding interactions. This suggests that inter-residue interactions are not entirely random in the thermally unfolded nuclease; similar suggestions have been made from NMR studies of other proteins⁶.

The significance of proline isomerism in protein folding has received considerable attention in recent years⁴. It is of interest, for example, to compare the case of ribonuclease A, in which there has been considerable controversy over this issue. Although some workers have claimed that ribonuclease can fold only with a full complement of native proline isomers⁷, others

have detected a metastable, native-like folding intermediate, which they interpret as containing a single proline in a non-native orientation^{8,9}. Our results with staphylococcal nuclease show that alternative proline isomers can indeed be accommodated in similar folded structures and that isomerization can occur in a folded protein molecule.

This work was supported by grants from SERC, NIH and NSF. P.A.E. acknowledges a Senior Hulme Scholarship from Brasenose College. C.M.D. is a member of the Oxford Enzyme Group.

Received 1 April; accepted 21 July 1987.

1. Fox, R. O., Evans, P. A. & Dobson, C. M. *Nature* **320**, 192-194 (1986).
2. Dobson, C. M. & Evans, P. A. *Biochemistry* **23**, 4267-4270 (1984).
3. Brandts, J. F., Halvorson, H. R. & Brennan, M. *Biochemistry* **14**, 4953-4963 (1975).
4. Kim, P. S. & Baldwin, R. L. *A. Rev. Biochem.* **51**, 459-489 (1982).
5. Arnone, A. *et al. J. biol. Chem.* **246**, 2302-2316 (1971).
6. Dobson, C. M., Evans, P. A. & Williamson, K. L. *FEBS Lett.* **168**, 331-334 (1984).
7. Lin, L. N. & Brandts, J. F. *Biochemistry* **23**, 5713-5723 (1984).
8. Schmid, F. X. & Baldwin, R. L. *Proc. natn. Acad. Sci. U.S.A.* **75**, 4764-4768 (1978).
9. Schmid, F. X., Grafl, R., Wrba, A. & Beintema, J. J. *Proc. natn. Acad. Sci. U.S.A.* **83**, 872-876 (1986).

Catalysis of protein folding by prolyl isomerase

Kurt Lang*, Franz X. Schmid*† & Gunter Fischer†

* Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Universitätsstraße 31, D-8400 Regensburg, FRG
† WB Biochemie, Sektion Biowissenschaften, Martin-Luther-Universität Halle Wittenberg, DDR-4020 Halle, GDR

Rates of protein folding reactions vary considerably. Some denatured proteins regain the native conformation within milliseconds or seconds, whereas others refold very slowly in the time range of minutes or hours. Varying folding rates are observed not only for different proteins, but can also be detected for single polypeptide species. This originates from the co-existence of fast- and slow-folding forms of the unfolded protein, which regain the native state with different rates¹⁻⁴. The proline hypothesis provides a plausible explanation for this heterogeneity. It assumes that the slow-folding molecules possess non-native isomers of peptide bonds between proline and another residue, and that crucial steps in the refolding of the slow-folding molecules are limited in rate by the slow reisomerization of such incorrect proline peptide bonds⁵⁻⁹. Recently the enzyme peptidyl-prolyl *cis-trans* isomerase (PPIase) was discovered and purified from pig kidney. It catalyses efficiently the *cis-trans* isomerization of proline imidic peptide bonds in oligopeptides^{10,11}. Here we show that it also catalyses slow steps in the refolding of a number of proteins of which fast- and slow-folding species have been observed and where it was suggested that proline isomerization was involved in slow refolding. The efficiency of catalysis depends on the accessibility for the isomerase of the particular proline peptide bonds in the refolding protein chain.

We first examined the effect of prolyl isomerase on the refolding of the immunoglobulin light chain from mouse. Fast- (U_F) and slow-folding (U_S) species have been found previously in unfolded light chain and isomerization of Pro 143 was thought to be responsible for the formation of U_S molecules in the unfolded state and for the low refolding rate of the U_S species³. The refolding experiments throughout this work were carried out at 10 °C in 0.1 M Tris-HCl, pH 8.0, in the presence of residual 0.25-0.30 M urea, the conditions for optimal activity

† To whom correspondence should be addressed.

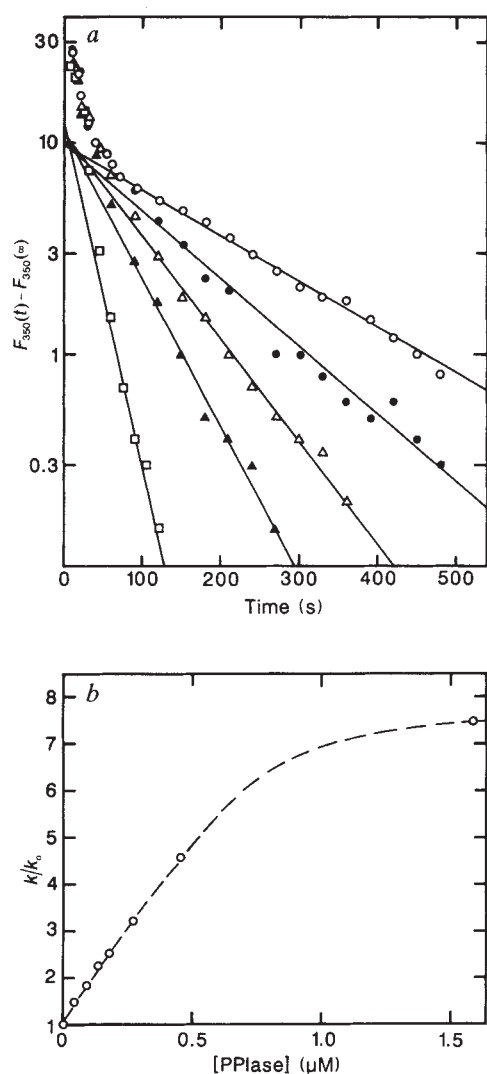


Fig. 1 *a*, Refolding kinetics of the immunoglobulin light chain in the presence of increasing amounts of PPIase. The decrease in fluorescence at 350 nm (F_{350}) is shown as a function of the time of refolding (t). The final conditions for refolding were 2 μM light chains in 0.25 M urea and 0.1 M Tris-HCl, pH 8.0 at 10 °C. The refolding solution contained the following concentrations of PPIase, $\circ$, 0 μM ; $\bullet$, 0.05 μM ; $\triangle$, 0.15 μM ; $\blacktriangle$, 0.3 μM ; $\square$, 1.6 μM . The slow phases of the observed kinetics were analysed in terms of first-order processes resulting in time constants (τ) of 202 s (0 μM PPIase), 136 s (0.05 μM), 89 s (0.15 μM), 63 s (0.3 μM) and 27 s (1.6 μM). The fast phase ($\tau = 15$ s) is not affected by the isomerase. *b*, Dependence on PPIase concentration of the relative increase of the apparent rate of slow refolding of the light chains. The ratios of the observed rate constants in the presence, k , and in the absence, k_0 , of PPIase are shown.

Methods. Murine immunoglobulin light chains (κ -type) were from Boehringer Mannheim. PPIase was isolated from pig kidney by a slightly modified version of the method described in ref. 10. Concentrations of PPIase were determined using a relative molecular mass (M_r) of 14,000 and a molar absorbance of $7,000 \text{ M}^{-1} \text{ cm}^{-1}$ (K.L., unpublished data). Isomerase activity was assayed as in ref. 10. The specific activity of the employed PPIase was 188 U mg^{-1} . Urea, p.a. grade and all other chemicals were from Merck. Unfolded protein was prepared by a 1 h incubation of 40 μM light chain in 5.0 M urea, 0.1 M Tris-HCl, pH 8.0 at 25 °C. Refolding was initiated by diluting 35 μl of the unfolded protein with 665 μl of 0.1 M Tris-HCl, pH 8.0 containing variable amounts of PPIase in the spectrometer cell at 10 °C. Refolding was monitored by the decrease in tryptophan fluorescence at 350 nm (10-nm slit). Excitation was at 280 nm (2-nm slit). A Hitachi Perkin-Elmer MPF 44A fluorescence spectrophotometer was used.

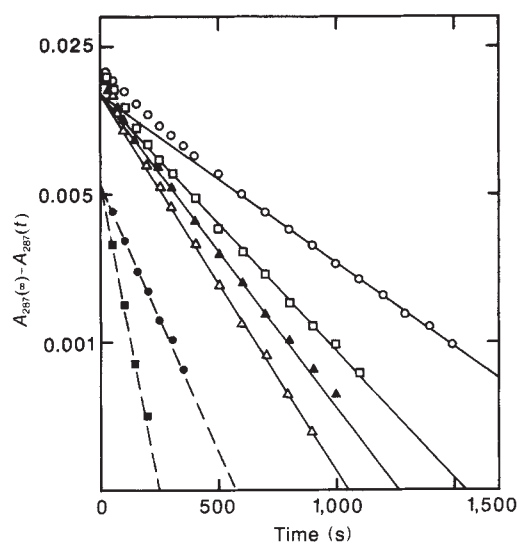


Fig. 2 Slow refolding kinetics of the S-protein fragment of bovine RNase A (17 μM) in 0.25 M urea, pH 8.0 at 10 °C in the absence ($\circ$) and in the presence of 3 μM ($\square$), 6 μM ($\blacktriangle$), and 9 μM ($\triangle$) PPIase during refolding. Refolding kinetics were analysed as the sum of two exponentials, as indicated by the lines for the slow phase. The observed time constants for the slow phase are 650 s (0 μM), 370 s (3 μM), 285 s (6 μM), and 245 s (9 μM PPIase). The corresponding data for the second slow phase are shown for 0 μM PPIase ($\bullet$, $\tau = 160$ s) and for 3 μM PPIase ($\blacksquare$, $\tau = 80$ s) only, to avoid crowding the figure.

Methods. S-protein was prepared from bovine RNase A (type XII A, Sigma) and purified as described¹⁷. S-protein (270 μM) was unfolded by a 1 h incubation in 4.0 M urea, 0.1 M glycine-HCl at pH 2.0 and 25 °C. Refolding was initiated by mixing 50 μl of unfolded protein with 750 μl of 0.1 M Tris-HCl pH 8.2 containing variable amounts of PPIase to give final conditions of pH 8.0, 10 °C and 17 μM S-protein. Kinetics were monitored by the increase in absorbance at 287 nm with a Cary 118C spectrometer. Control experiments in the presence of heat-inactivated PPIase or bovine serum albumin did not result in an acceleration of the refolding kinetics.

of PPIase. Small concentrations of guanidinium chloride and pH values below 7 were found to decrease the isomerase activity substantially. Refolding was initiated by a 20-fold dilution of unfolded protein (in 5 M urea, pH 8.0) with the refolding solution with variable amounts of PPIase. In the absence of isomerase, refolding of the light chain is biphasic³: a fast phase ($\tau = 15$ s) accounts for about 75% of the total amplitude and a slow phase ($\tau = 200$ s) for 25%. Figure 1a shows the effect of increasing concentrations of PPIase on the refolding kinetics of the light chain. The fast kinetic phase is unaffected by PPIase, but the slow refolding reaction is catalysed by it. The extent of acceleration depends on the concentration of PPIase. The relative increase in folding rate, the ratio of the apparent rate constants in the presence and in the absence of PPIase, is shown in Fig. 1b as a function of the isomerase concentration. PPIase at 0.1 μM (this corresponds to a molar ratio PPIase/light chain of 1:20) leads to a twofold increase of the rate of slow refolding. In high concentrations of PPIase, the slow refolding reaction coincides with the fast phase.

Fast- and slow-refolding species were first detected for bovine RNase A (ref. 1). Subsequently this protein served as a model system for elucidating the function of proline isomerization in the formation of U_s species in the unfolded polypeptide chain and its importance for rate-limiting slow steps in the course of refolding^{4-9,12}. Therefore we investigated the effect of PPIase on the slow refolding of two different RNases with varying proline content and on the folding of a fragment of RNase A, the S-protein.

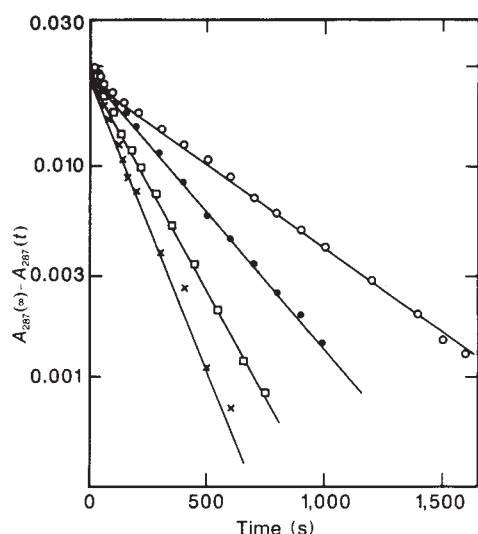


Fig. 3 Refolding kinetics of porcine pancreatic RNase (11 μ M) in 0.30 M urea, pH 8.0, 10 °C in the absence (○, τ = 540 s) and in the presence of 1.5 μ M (●, τ = 330 s), 3.0 μ M (□, τ = 220 s) and 4.5 μ M (×, τ = 165 s) PPIase during refolding. Refolding was carried out by a 20-fold dilution of unfolded RNase (220 μ M in 6.0 M urea, 0.1 M glycine-HCl, pH 2.0) with 0.1 M Tris-HCl, pH 8.2, 10 °C and the indicated concentrations of PPIase. Porcine RNase was isolated as described¹⁶ and refolding kinetics were determined as in Fig. 2.

The slow refolding reaction of bovine RNase A at 10 °C, pH 8.0, in 0.25 M urea, consists of two parallel processes: a major slow phase (about 65% of all unfolded molecules) and a minor very slow phase (about 15%). The fast refolding reaction (20%) takes place in less than a second. The major slow phase involves the rapid formation of an intermediate with a native-like structure and a subsequent slow step, which, it has been suggested, is *trans* $\rightarrow$ *cis* isomerization of Pro 93 of bovine RNase A (refs 7 and 12). This major slow refolding reaction cannot be accelerated by addition of prolyl isomerase, confirming earlier results¹¹. Catalysis of the minor, very slow, phase of refolding is uncertain. Because of its small amplitude, this reaction and the influence of PPIase on its rate are difficult to measure.

Removal of 20 residues from the amino terminus of RNase A by subtilisin and separation of the 1–20 peptide leads to the S-protein fragment (residues 21–124) (ref. 13). The S-protein is much less stable than intact RNase A, but it contains all four proline residues of the parent protein and it displays the same distribution of U_F and U_S species as intact RNase A (ref. 4). Refolding of U_S of S-protein is very slow and partially structured intermediates could not be detected in the course of slow refolding, unlike the folding of RNase A. Figure 2 shows that the entire slow phase of S-protein refolding is catalysed by the isomerase. PPIase at $\sim 3 \mu$ M is required for a twofold acceleration of both slow folding reactions. This corresponds to a 1:5 ratio of isomerase to S-protein. Heat-inactivated PPIase or equivalent amounts of bovine serum albumin had no effect on the refolding rates. The comparison of the results obtained with intact RNase A and with the S-protein fragment suggests that proline isomerization is involved in the entire slow phase of refolding. In the case of intact RNase A, partially folded intermediates are formed rapidly and render one or more prolines inaccessible for the isomerase. Consequently slow refolding cannot be catalysed by PPIase. The S-protein is only marginally stable and partially folded intermediates with incorrect proline isomers are not populated during refolding. Therefore the prolines remain accessible and folding is catalysed by PPIase.

Porcine RNase differs from other pancreatic RNases by an additional proline residue at position 115, thus leading to a Pro 114–Pro 115 sequence¹⁵. Isomerization of this sequence in the

unfolded molecule is about 7-fold slower than the isomerization of other prolines in RNases and it leads to an additional unfolded form ($\sim 70\%$ of all molecules) which refolds very slowly¹⁶. Figure 3 shows that the extra-slow refolding reaction of porcine RNase is also catalysed by PPIase. This corroborates the suggestion that the difference in folding mechanism between porcine and bovine RNases originates from a proline substitution, the replacement of Tyr 115 by proline in the porcine enzyme¹⁶. It also indicates that the isomerase accepts Pro–Pro-containing chain segments as a substrate.

To summarize, we have shown for a number of proteins that slow steps in *in vitro* refolding can be catalysed by peptidyl-prolyl *cis-trans* isomerase. This enzyme efficiently accelerates the *cis* $\rightleftharpoons$ *trans* isomerization of proline peptide bonds in oligopeptides¹⁰. Other catalytic activities have not yet been detected. Our results indicate that proline isomerization is indeed the rate-limiting molecular event in a number of slow protein folding reactions. The accessibility of the particular proline peptide bonds for the isomerase is important for efficient catalysis during refolding. Structure formation in proline-containing chain segments may render them inaccessible for the isomerase. The results obtained for the various RNase species support the suggestion that the entire slow refolding reactions of these proteins involve proline isomerization. Processes other than proline isomerization⁹ are probably not required to explain the existence of slow-folding molecules.

Received 8 June, accepted 23 July 1987.

1. Garel, J.-R. & Baldwin, R. L. *Proc. natn. Acad. Sci. U.S.A.* **70**, 3347–3351 (1973).
2. Kato, S. *et al. Biochemistry* **21**, 38–43 (1982).
3. Goto, Y. & Hamaguchi, K. *J. molec. Biol.* **156**, 891–910 (1982).
4. Kim, P. S. & Baldwin, R. L. *A. Rev. Biochem.* **51**, 459–489 (1982).
5. Brandts, J. F. *et al. Biochemistry* **14**, 4953–4963 (1975).
6. Schmid, F. X. & Baldwin, R. L. *Proc. natn. Acad. Sci. U.S.A.* **75**, 4764–4768 (1978).
7. Schmid, F. X. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 872–876 (1986).
8. Lin, L.-N. & Brandts, J. F. *Biochemistry* **22**, 573–580 (1983).
9. Lin, L.-N. & Brandts, J. F. *Biochemistry* **23**, 5713–5723 (1984).
10. Fischer, G. *et al. Biomed. Biochim. Acta* **43**, 1101–1111 (1984).
11. Fischer, G. & Bang, H. *Biochim. biophys. Acta* **828**, 39–42 (1985).
12. Schmid, F. X. & Blaschek, H. *Eur. J. Biochem.* **114**, 111–117 (1981).
13. Richards, F. M. & Vithayathil, P. J. *J. biol. Chem.* **234**, 1459–1465 (1959).
14. Labhardt, A. M. & Baldwin, R. L. *J. molec. Biol.* **135**, 231–244 (1979).
15. Jackson, R. L. & Hirs, C. H. W. *J. biol. Chem.* **245**, 637–653 (1970).
16. Grafl, R. *et al. J. molec. Biol.* **191**, 281–293 (1986).
17. Doscher, M. S. & Hirs, C. H. W. *Biochemistry* **6**, 304–312 (1967).

Conformational mutations in human mitochondrial DNA

Gurparkash Singh, Nicolas Neckelmann
& Douglas C. Wallace*

Departments of Biochemistry, Pediatrics and Anthropology,
Emory University School of Medicine, Woodruff Memorial Building,
Atlanta, Georgia 30322, USA

Variation in the human mitochondrial DNA (mtDNA) sequence has been extensively analysed using restriction fragment length polymorphisms (RFLPs)^{1–11}. MtDNA RFLPs have previously been attributed to nucleotide changes within restriction endonuclease recognition sites^{1–10} or to small insertion–deletion mutations¹¹. We now report that RFLPs detected by polyacrylamide gel electrophoresis can also result from single nucleotide substitutions which alter the mobility of small- to medium-sized restriction fragments that incorporate the sequence. We have defined the mutation responsible at two loci and have identified several possible additional loci. When screening human mtDNAs with multiple restriction endonucleases, such mutations can be misidentified as insertion–deletion mutations or counted as multiple polymorphic restriction sites. This can lead to errors in constructing restriction maps and estimating sequence diversity.

* To whom correspondence should be addressed.

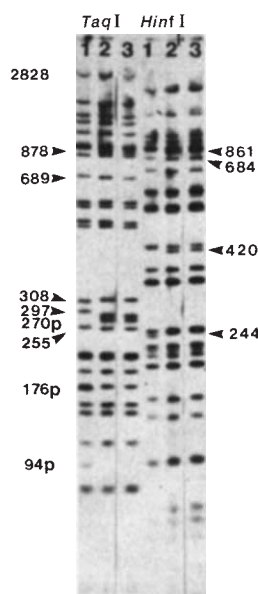


Fig. 1 Autoradiograph of *TaqI* and *HinfI* restriction fragments of HeLa (lane 1), HT1080 (lane 2), and GM6224 (lane 3) mtDNA resolved on a 3.5% polyacrylamide gel. The variant mobility fragments are indicated by arrows. In HeLa cells, the 270-np *TaqI* fragment (270p) contains an additional polymorphic site generating 176p and 94p fragments²⁴. Restriction mapping of mobility variants: The 689- and 878-np *TaqI* and 684- and 861-np *HinfI* fragments were identified by double digestion; 255- and 297-np *TaqI* fragments by sequence analysis and the remainder of variant fragments by size estimation relative to standards. The 255-np *TaqI* fragment (at 10,180–10,435 nps) maps to the same region as the 684-np *HinfI* fragment (at 9,753–10,437 nps). The 420-np *HinfI* fragment (at 10,971–11,391) overlaps 297 (at 11,124–11,421) and 689-np *TaqI* fragments (10,435–11,124). The 878-np *TaqI* fragment (12,526–13,404) correlates with the 861-np *HinfI* fragment (12,170–13,031). The 244-np *HinfI* fragment (15,756–16,000) is contained in the 2,828-np *TaqI* fragment (14,956–1,215) and the 308-np *TaqI* fragment (9,443–9,751) is encompassed by a 970-np *HinfI* fragment (8,783–9,753). The 970-np *HinfI* and 2,828-np *TaqI* fragments are presumably too large to be affected by mobility mutations.

Methods. Mitochondria were isolated from homogenized cultured cells by differential centrifugation and the mtDNAs extracted by SDS-lysis and purified by isopycnic CsCl-ethidium bromide density gradient centrifugation⁴. The mtDNAs were digested with restriction endonucleases according to the manufacturers' instruction and the restriction fragments were end-labelled with [α -³²]dATP and resolved in polyacrylamide gels². The 297 and 255-np *TaqI* fragments were cloned into the *AccI* site of M13 vectors and sequenced by the dideoxy chain termination method²⁷.

Human mtDNA is a closed circular molecule having 16,569 nucleotide pairs (np)¹². While analysing mtDNA RFLPs on polyacrylamide gels², we detected small deviations in fragment mobility in protein coding regions similar to those previously attributed to small insertion-deletion mutations in non-coding regions¹¹. Some of the electrophoretic mobility variants observed are shown in Fig. 1. *TaqI* and *HinfI* restriction fragments of mtDNA from three human cell lines (HeLa, HT1080 and GM6224) were end-labelled and resolved on a polyacrylamide gel². Mobilities of fragments of HT1080 and GM6224 mtDNAs are the same, but differ from HeLa mtDNA in 5 of 29 *TaqI* fragments of which the HeLa 255, 297, 689 and 878-np mtDNA fragments move more slowly than those from HT1080 and GM6224 while the HeLa 308-np mtDNA fragment moves faster.

Four of the 36 *HinfI* fragments (244, 420, 684 and 861 np) also have altered electrophoretic mobility, the HeLa 420 and

861-np fragments migrate more slowly than HT1080 and GM6224, and the HeLa 244-np fragment faster. The 684-np *HinfI* fragment is retarded in all three cell lines with respect to external standards and *TaqI* fragments (Fig. 1). These differences in mobility are consistently discernible in a range of gel concentrations (3.5%, 8.0% and 12%).

Several of the variant *TaqI* and *HinfI* fragments map to the same mtDNA region (such as the 297-np *TaqI* and the 420-np *HinfI* fragments, Fig. 1) suggesting that they are the product of common sequence changes.

To determine the molecular basis of the variations in mobility, we cloned and sequenced two of the variant *TaqI* fragments (255 and 297 np) from four different human cell lines (HeLa, HT1080, LM, GM6224 and KN). The sequence differences are presented in Table 1. Surprisingly, all of the fragments were exactly the same length. Hence, the altered mobility is not the

Table 1 Sequence differences of variant mobility fragments

<i>TaqI</i> fragment position	mtDNA origin	Fragment size (np)	Relative mobility	Position of variant nucleotide		
11,124–11,421	Cam*	297	NT	11,253	11,335	11,347
	HeLa	297	S	T	T	A
	HT1080	297	N	T	C	G
	GM6224	297	N	C	C	A
	LM	297	S	C	C	A
10,180–10,435	Cam*	255	NT	10,398	10,373	10,370
	HeLa†	255	S	A	G	T
	HT1080	255	N	G	A	T
	LM	255	N	A	G	T
	KN	255	S	G	G	C
		255	N	A	G	T

* Cam is published sequence¹², all np positions correspond to Cam numbering.

† A polymorphic base at 10,259 reported in Oliver *et al.*²⁴ was not observed in this study.

‡ NT, not tested; N, normal; S, slow.

§ For the 297 np fragments we have sequenced 3 independent HeLa clones, 2 HT1080, 2 GM6224 and 2 LM. For the 255-np fragments we have sequenced 2 independent HeLa clones, 1 HT1080, 4 LM and 1 KN. We have confirmed that the relative mobility differences are retained after cloning by releasing the 297-np inserts of HeLa, HT1080 and GM6224 from M13 RFs, end-labelling the fragments, and resolving the fragments on polyacrylamide gels.

|| Indicates replacement mutations, 11,253 is Ile (Cam) to Thr (HT1080) and 10,398 is Thr (Cam) to Ala (HeLa).

Table 2 List of inferred restriction sites proposed by Cann and co-workers²³

Position enzyme*	Site		RFLP reported Cam† → HeLa†	Site† configuration
	present (+) or absent (-) in Cam†	HeLa†		
10,066 d	-	+	2311 → 686 + 1625	A
10,084 f	-	+	270 → 84 + 176	P
10,352 a	-	+	366 → 120 + 246	A
10,394 b	-	+	275 → 38 + 237	P
11,146 b	-	+	695 → 515 + 180	A
11,350 a	+	-	440 + 12 → 452	A

* The symbols a, b, c, d, e and f refer to restriction endonuclease sites *AluI*, *DdeI*, *HaeIII*, *HhaI*, *HpaII* and *TaqI*, respectively. Additional polymorphic sites were reported at 10,364c, 10,413a (10,536), 11,161e and 11,329c (11,690) nps. These were listed as the same for HeLa and Cam and could not be checked using our data. Parentheses refer to alternate placement of inferred sites^{3,23}.

† Cam is a published sequence¹²; Cam and HeLa are samples 110 and 45, respectively in Cann *et al.*²³.

‡ A, site absent in sequence; P, site present in sequence. The sequence surrounding the HeLa-specific *TaqI* site at 10,084f was reported by us previously²⁴.

result of either restriction site polymorphisms at the ends of fragments^{2,3} or insertion-deletion mutations¹¹. The sequences do differ by several transition mutations. In the 297-np fragment there are three polymorphic nucleotides (at 11,253, 11,335 and 11,347 nps) of which only the variant at 11,253 np gives rise to differences in electrophoretic migration. In case of the HeLa and LM fragments a thymine substitution correlates with a reduced mobility compared to the HT1080 and GM6224 fragments.

There are also three polymorphic sites in the 255 np-fragment (at positions 10,398, 10,373 and 10,370 nps). The slower HeLa fragment differs from the faster HT1080 and KN fragments at 10,398 and 10,373 nps, and the slower LM fragment differs from HT1080 and KN at 10,398 and 10,370 nps. Only a G residue at nucleotide 10,398 correlates with reduced mobility. Therefore, base substitutions in human mtDNA can change electrophoretic mobilities of restriction fragments.

The most reasonable explanation for the altered mobility of these fragments is that the mutations affect the conformation of the DNA. Changes in restriction fragment mobilities attributed to sequence-specific variations in DNA conformations have been reported for *Salmonella*¹³, yeast¹⁴, bacteriophage λ ¹⁵ and trypanosomatid minicircle DNAs^{16,17} in which 'bent' DNA alters the migration of restriction fragments in polyacrylamide gels but not in agarose gels¹⁶.

It has been proposed that bending of DNA results from the periodic disposition of ApA (TpT) dinucleotides^{18,19}, or from

the non-equivalence of DNA helices at A-tract junctions^{20,21}. Either model would predict a conformation change in the 297-np *TaqI* fragment. For example, application of the ApA model to the 10-np sequence (NTAATTTANN) of HeLa surrounding polymorphic nucleotide at position 11,253 nps, repeated 20 times, gives a circle of a circumference 187 np. This estimate is comparable to the 155 np predicted for the ideal 'bent' sequence (NNNNAAAAAA)²¹. The allelic sequence (NTAACTTANN) of HT1080 gives a circle of circumference ~310 np. The DNA bending of the HeLa 297 np-fragment is therefore consistent with a reduced mobility.

The biological significance of mtDNA conformation mutants remains uncertain, but it is clear that when RFLPs are being analysed on polyacrylamide gels, mobility mutants can be misinterpreted as insertion-deletion mutations or as restriction site polymorphisms. Confusion between conformation and length mutations could occur in the region of our 255-np variant which maps to the same position as the 'insertion-deletion' length variant region VI reported by Cann and Wilson¹¹, who assigned a 10-np insertion to this region between the ND3 and tRNA^{arg} genes without sequencing it. Mobility variants in the mtDNA could therefore result from length mutations²² or from conformational mutations.

Confusion between conformation mutations and restriction site polymorphisms is possible in the region of both the 255- and 297-np fragments^{2,3,23}. Cann *et al.* have assigned 10 polymorphic restriction sites in the regions 10,014–10,830²⁴ and 11,124–11,424 nps, of which six are reported to be polymorphic between HeLa (their sample 45) and the published sequence¹² (their sample 110)²³. But comparison of our HeLa sequences with the published sequence (Table 1) failed to confirm four of the six polymorphisms (Table 2): the first of these (site 10,066d) appears to be a mapping error, and the other three (10,352a, 11,146b and 11,350a) could be the product of mobility variant point mutations. It is possible that additional mobility variants have been attributed to 'restriction site polymorphisms' in human mtDNA. Cann *et al.* inferred from their RFLPs that the ratio of transition to transversion mutations was 2.5:1 (ref. 3) but sequence data suggest that the ratio is as high as 25:1 (ref. 25). Random selection of restriction sites should give a ratio of 1:2.

Variant mobility fragments have been reported in human¹¹, sea urchin²⁶ and mouse mtDNAs^{28,29}. If they are a general feature of mtDNAs, their detection by polyacrylamide gel and inclusion in RFLP data could bias calculations of mutation distances and distort mtDNA phylogenies²³. Conformational mutations are not detected on agarose gels¹⁶ and consequently they have not been a source of error in analyses of human mtDNA RFLPs^{4–10} employing this method.

We thank Dr D. M. Crothers for help with computer prediction of DNA conformation changes and Tad Schurr for sequence analysis. This work was supported by grants from the NSF, NIH, American Cancer Society, and Muscular Dystrophy Foundation (D.C.W.).

Received 18 May; accepted 30 July 1987.

- Brown, W. M. & Goodman, H. M. in *Extrachromosomal DNA* (eds Cummings, D. J. *et al.*) 15, 485–499 (Academic, New York, 1979).
- Brown, W. M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3605–3609 (1980).
- Cann, R. L., Brown, W. M. & Wilson, A. C. *Genetics* **106**, 479–499 (1984).
- Denaro, M. *et al. Proc. natn. Acad. Sci. U.S.A.* **78**, 5768–5772 (1981).
- Blanc, H., Chen, K.-H., D'Amore, M. A. & Wallace, D. C. *Am. J. hum. Genet.* **35**, 167–176 (1983).
- Johnson, M. J., Wallace, D. C., Ferris, S. D., Rattazzi, M. C. & Cavalli-Forsza, L. L. *J. molec. Evol.* **19**, 255–271 (1983).
- Wallace, D. C., Garrison, K. & Knowler, W. C. *Am. J. phys. Anthropol.* **68**, 149–155 (1985).
- Brega, A. *et al. Am. J. hum. Genet.* **39**, 502–512 (1986).
- Brega, A. *et al. Ann. hum. Genet.* **50**, 327–338 (1986).
- Wallace, D. C. in *Medical and Experimental Mammalian Genetics: A Perspective* (eds McKusick, V. *et al.*) 137–190 (Liss, New York, 1987).
- Cann, R. L. & Wilson, A. C. *Genetics* **104**, 699–711 (1983).
- Anderson, S. *et al. Nature* **290**, 457–465 (1981).

- Bossi, L. & Smith, D. M. *Cell* **39**, 643–652 (1984).
- Snyder, M., Buchman, A. R. & Davis, R. W. *Nature* **324**, 87–89 (1986).
- Zahn, K. & Blattner, R. F. *Nature* **317**, 451–453 (1985).
- Kitchin, P. A. *et al. J. biol. Chem.* **261**, 11302–11309 (1986).
- Griffith, J., Bleyman, M., Rauch, C. A., Kitchin, P. A. & Englund, P. T. *Cell* **46**, 717–724 (1986).
- Trifonov, E. N., & Sussman, J. L. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3816–3820 (1980).
- Ulanovsky, L. E. & Trifonov, E. N. *Nature* **326**, 720–722 (1987).
- Wu, H.-M. & Crothers, D. M. *Nature* **308**, 509–513 (1984).
- Koo, H.-S., Wu, H.-M. & Crothers, D. M. *Nature* **320**, 501–506 (1986).
- Wrischnik, L. A. *et al. Nucleic Acids Res.* **15**, 529–542 (1987).
- Cann, R. L., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
- Oliver, N. A., Greenberg, B. D. & Wallace, D. C. *J. biol. Chem.* **259**, 5834–5839 (1983).
- Greenberg, B. D., Newbold, J. E. & Sugino, A. *Gene* **21**, 33–49 (1983).
- Vawter, L. & Brown, W. M. *Science* **234**, 194–196 (1986).
- Sanger, F., Nicklens, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).
- Howell, N. *Plasmid* **14**, 93–96 (1985).
- Howell, N., Appel, J., Cook, J. P., Howell, B. & Hauswirth, W. W. *J. biol. Chem.* **262**, 2411–2414 (1987).